2008

Integrating QTL analysis into plant breeding practice using Bayesian statistics

Shengqiang Zhong
Iowa State University

Integrating QTL analysis into plant breeding practice using Bayesian statistics

by

Shengqiang Zhong

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Jean-Luc Jannink, Co-major Professor
Jack C. M. Dekkers, Co-major Professor
Alicia L. Carriquiry
Michael Lee
Rohan L. Fernando
Jode W. Edwards

Iowa State University

Ames, Iowa

2008

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I want to take this opportunity to thank all people who helped me during my Ph.D. study. It is not possible for me to list all the names here and I apologize in advance if omit someone's name.

My extremely grateful thanks go to my advisors Dr. Jean-Luc Jannink and Dr. Jack Dekkers for their guidance, patience and support. Dr. Jannink was very understanding and always available for great instruction, and encouragement throughout my Ph.D. study. My whole thesis has benefited enormously from Dr. Jannink's sharp insight in scientific research and from his extensive knowledge in quantitative genetics. Dr. Dekkers provided a new perspective and invaluable suggestion to my study. I would like to thank other committee members Dr. Alicia L. Carriquiry, Dr. Michael Lee, Dr. Rohan L. Fernando and Dr. Jode W. Edwards for their advice and support from various aspects.

I want to thank the students in the Jannink's Lab: Murli Gogula, Jin Long, Alona Chernyshova, Yoon-Soup So, Massiel Orellana, and Julia Olmstead and Lucía Gutiérrez for all their support and sharing the fun time in the field. I am thankful to George Patrick and Ron Skrdla from Jannink's group for sharing field breeding experience. I also want to thank the researchers from the genomic selection group in the Animal Science Department for sharing their experience.

I would like to express my appreciation to all the faculty members, students and staffs in the Agronomy Department for providing me with such a professional research environment.

Finally, I want to particularly thank my girlfriend, my parents, my girlfriend's parents and my brother and sister. It is their support and love that make all what I have accomplished so far possible.

ABSTRACT

Plant breeders face the challenges to incorporate significant developments in Bayesian quantitative trait locus (QTL) analysis into breeding practice. The overall objective of this dissertation is to integrate QTL analysis into marker-assisted selection (MAS) in plant breeding using Bayesian statistics. Three different perspectives were studied: identification of optimal designs to generate multiple families for QTL mapping, cross prediction using QTL analysis results, and genomic selection for cultivated barley. First, the impact of two mating designs was studied on QTL mapping in multiple families generated by crosses between multiple inbred lines, using within-family linkage disequilibrium (LD) with a Bayesian variable selection approach. The loop design was found to have smaller mean square error in estimating QTL allelic variance and position. Second, the usefulness of crosses in developing inbred lines was investigated, using QTL mapping results from Bayesian shrinkage analysis. The usefulness of a particular cross depends on the expected performance of its best progeny, which was called the superior progeny value here. Theory was developed to predict the superior progeny value as a function of the mean of the breeding values of all progeny and of the standard deviation of the breeding values among progeny from a specific cross. Little difference among crosses for the standard deviation among their progeny was found under an additive genetic model for a trait, such that a benefit from estimating that standard deviation occurred only in relatively few cases. Finally barley marker data from 1803 SNP was used to evaluate genomic selection for breeding populations derived from 42 spring two-row barley lines. Three different genomic selection methods, random regression best linear unbiased prediction (RR-BLUP), Bayesian shrinkage estimation, and Bayes-B, were compared. The current barley SNP density was found to be high enough for genomic selection to better predict the breeding values of double haploid progeny than phenotypic selection. Overall, the Bayes-B approach that fitted a relatively high proportion of markers into the model had more stable performance across different scenarios. MAS for quantitative traits in plant breeding seem promising by integrating advanced Bayesian QTL analysis.

CHAPTER I. INTRODUCTION

Traditional plant breeding has depended on phenotypic selection for agronomically important traits. Significant crop improvements through phenotypic selection have been made (Fehr, 1984). However phenotypic selection in some situations presents difficulties due to genotype-environment interactions and unreliable, or expensive phenotyping. With recent considerable marker technology developments, molecular marker-assisted selection (MAS), that is, selection for the genetic determinant or determinants of a trait of interest based on marker genotypes, offers a great opportunity for efficient selection for traits controlled both by major genes as well as by many quantitative trait loci (QTL). In this chapter, I first briefly review genetic markers, MAS for qualitative traits, and MAS for quantitative traits. Then I place the research described in this thesis against this background.

Genetic Markers

Genetic markers can be classified into three major groups: phenotypic markers, biochemical markers and DNA sequence based markers. Phenotypic markers are generally visually characterized morphological characters, such as pea flower color and seed shape determined by Mendelian genes. Some phenotypic markers are easy to score and if they are linked to important agricultural traits such as disease resistance, they can be used in breeding programs (Joshi et al 2004). Biochemical markers, mostly isozyme markers, usually exploit different variants of the same enzyme (Weeden and Wendel, 1990). Phenotypic and biochemical markers are limited in number and might be subjected to variation due to environment or developmental stage, so their application in breeding is restricted (Winter and Kahl, 1995).

Markers based on DNA sequence, which we will call simply molecular markers, are the most widely used type of marker, due to their abundance and their independence of environment and development. Since the 1980's, different types of molecular markers have been developed, including restriction fragment length polymorphisms (RFLPs), random amplification of polymorphic DNA (RAPDs), amplified fragment length polymorphisms (AFLPs), simple sequence repeats (SSRs) or microsatellite markers, and single nucleotide

polymorphisms (SNPs). The RFLP is the oldest type of DNA marker. Large amounts of DNA are typically required for RFLP detection and it is difficult to automate the analysis (Erlich and Arnheim, 1992). Other types of molecular markers generally require smaller amounts of DNA. The RAPD, which utilizes single primers of arbitrary sequence to generate strain-specific arrays of anonymous DNA fragments by low stringency PCR amplification (Wang et al 1993), is relatively unreliable in terms of reproducibility. The AFLP requires DNA digestion by restriction enzymes, then using PCR with selective primers to amplify specific fragments (Vos and Zabeau, 1993). The SSR is based on short-repeat sequences that are widely dispersed throughout the eukaryotic genome. SSRs are highly informative because their mutation mechanism generates variable numbers of repeats, and thus many different marker alleles are possible. A SNP marker is a difference in nucleotide between different alleles, at a single base pair position in the genome. Due to the abundance of SNPs and the development of sophisticated high-throughput of SNP detection systems, SNP usage has increased in QTL mapping and MAS. Continuously decreasing SNP cost will likely remove this important factor that limits MAS implementation.

The development of abundant DNA markers has made many marker applications possible: estimation of genetic diversity (Smith et al 2000), germplasm characterization (Mason et al 2005), construction of linkage maps (Luo et al 2001), qualitative gene and QTL mapping, the discovery of useful candidate genes (Thornsberry et al 2001; Blair et al 2003), and MAS (Brahm et al 2000; Willcox et al. 2002).

For breeding purposes, molecular markers can be applied in several ways:

1) Marker-assisted backcrossing (Willcox et al. 2002).
2) Gene pyramiding (Servin et al. 2004).
3) Selecting superior individuals within a population based on marker-estimated breeding values (Lande and Thompson 1990; Meuwissen et al. 2001).
4) Selecting best crosses among a set of lines (Zhong and Jannink, 2007).

Most markers are not the causal mutations themselves but are useful because they are linked to them. The success of QTL mapping and MAS relies on the extent of linkage disequilibrium (LD) in the population between the markers and the causative genes of the trait. Linkage disequilibrium is the non-random association of alleles at different loci

(Lewontin and Kojima, 1960). Linkage disequilibrium in the population is mainly generated by mutation, selection, drift and migration, and dissipated by recombination (Falconer and Mackay, 1997). Existent LD will remain for many generations between tightly linked loci and decay over few generations for loosely linked or unlinked loci. Three kinds of markers can be distinguished based on LD of the markers with the loci that contribute to genetic variation for the trait in the population (Dekkers, 2004):

i) The molecular marker itself is the functional polymorphism, which is the most favorable situation for MAS. In this case, it could be ideally referred to as gene-assisted selection. While this kind of relationship is the most preferred one, it is also difficult to find this kind of markers.

ii) The marker is in LD with the functional mutation throughout the population. Population-wide LD can usually be found when markers and genes of interest are physically close to each other. Selection using these markers can be called LD-MAS.

iii) The marker is in linkage equilibrium (LE) with the functional mutation across the population. This is the most difficult and challenging situation for QTL mapping and MAS. Although a marker and a linked QTL may be in LE across the population, within a family LD will always exist, even between loosely linked loci. Within-family LD can be used to detect QTL and for MAS (Fernando and Grossman, 1989).

Marker Assisted Selection for Qualitative Traits

Marker-assisted Backcrossing

For the introgression of qualitative traits such as disease resistance, which are typically controlled by single genes, backcross breeding has been used for a long time (Allard, 1960). Marker-assisted selection has been routinely employed to assist backcross introgression of major genes into elite cultivars and to select alleles with major effects on high-value traits (Chen et al. 2000; Singh et al. 2001). Flanking markers around a target gene are used to track the desirable alleles in foreground selection, while markers dispersed throughout the genome are used to recover the recipient genotype in background selection (Hospital and Charcosset, 1997).

In conventional backcrossing programs using phenotypes, a minimum of five or six-backcrossing generations is required to transfer the desired allele and recover the recurrent background. Furthermore there is a risk linkage drag around the target gene, that is, that a large segment of the donor-parent genome will remain intact around the target gene. The expected proportion of the recurrent parent background genome is $1 - (1/2)^{t+1}$ after backcrossing for t generations. However, any specific backcross progeny will vary from this expectation due to chance. Background markers can identify progeny more similar to the recurrent parent and thus accelerate the recovery of recurrent parent genotype. Frisch et al (1998) used simulations to compare several different backcrossing strategies in terms of how quickly they recovered a large proportion of the recurrent parent genotype. They recommended a four-step sampling strategy to quickly recover the recurrent parent genotype, which includes: (1) selecting individuals carrying the target allele; (2) selecting individuals homozygous for the recurrent parent alleles at loci flanking the target locus; (3) selecting individuals homozygous for recurrent parent alleles at remaining loci on the same chromosome as the target allele; and (4) selecting one individual that is homozygous for recurrent parent alleles at most loci (across whole genome) among those that remain. They found that using this strategy one could expect the recovery of at least 96% of the recurrent parent genotype with 90% probability after three generations of backcrossing with a reasonable population size (50-100).

Chen et al. (2000) demonstrated a very good example of marker-assisted backcrossing. They improved the bacterial blight resistance of an elite rice line by introgressing Xa21, a wide-spectrum bacterial blight resistance gene, into this line by molecular marker-assisted backcrossing. They selected for target alleles at two markers tightly linked to Xa21 and for recurrent parent alleles at flanking markers outside of the gene region to decrease linkage drag during three backcross generations. In the third backcross generation, they used 128 RFLP markers for background selection and recovered a line that was essentially identical to the recurrent parent cultivar, but possessing the Xa21 allele.

Gene Pyramiding

Another MAS application for simple inherited traits is gene pyramiding, which involves multiple crosses between several parents (Servin et al 2004). Gene pyramiding can be applied

to enhance resistance to disease and insects by selecting for two or more than two genes at a time. The advantage of using markers in this case is that it allows the breeder to select for QTL-alleles that have same phenotypic effect, which can be nearly impossible with conventional breeding approaches. For example, due to the broad spectrum of blast resistance of Xa21 allele, a breeding line with Xa21 only cannot be distinguished from a breeding line with Xa21 and some other genes with similar resistance by conventional phenotypic approach. Using MAS, Singh et al (2001) pyramided three bacterial blast resistance genes (Xa5, Xa13 and Xa21) into an indica rice cultivar. Pyramiding of several resistance genes by marker-assisted breeding may lead to more durable resistance.

Marker Assisted Selection for Quantitative Traits

Lande and Thompson (1990) showed in their original paper how DNA marker information could improve estimates of breeding values for quantitative traits. Their essential requirement was that the observable markers be in LD with the unobservable QTL affecting the trait. Many important agricultural traits, such as yield, are under polygenic control with gene interactions (epistasis), strong environmental influence, and genotype-by-environment interaction on trait expression. Although MAS has been widely implemented for traits controlled by major genes using introgression and gene pyramiding, successful application of MAS to quantitative traits has been limited.

Limitations of Current MAS

Current MAS for quantitative traits in plant breeding has been constrained essentially by two factors. First, the statistical methods have taken marker effects to be fixed rather than random. As we will review below, this approach means that the effects of some causal loci will not be accounted for (the loci remain undetected), and the effects of other loci will be estimated with bias. Effective use of marker effects under a fixed model would require large amounts of accurate phenotypic data. With such data, however, phenotypic selection is effective itself, such that the gain from marker information is limited. Second, biotechnologies have until recently only allowed fairly low marker densities, requiring that markers and QTL be in fairly long-range LD. This requirement, in turn, has meant that the LD used was within families generated by bi-parental crosses, such that linkage phase

between the marker and the QTL would generally not be consistent from family to family. The consequence of inconsistency of marker-QTL linkage phase is that estimated marker allele effects would be only nested within family and meaningless across families. Fernando and Grossman (1989) developed QTL mapping and MAS with LE markers by modeling the co-segregation of markers and QTL within pedigree or family. Research in overcoming this second limiting factor (lack of marker density) is not the subject presented here. Rather, we simply take for granted that markers at high density are available such that markers exist that are in population-wide LD with QTL. In that case, marker-QTL phases are consistent across individuals and it becomes reasonable to estimate marker allele effects. In other words, we assume now that we are proceeding with LD-MAS in the sense of Dekkers (2004).

Statistical Developments of QTL Analysis

The primary problem of QTL analysis with high-density marker data is that the number of independent variables (i.e., the number of markers) is large relative to the number of observations (i.e., the number of genotypes observed). Several variable selection strategies have been developed and applied to address this problem. Step-wise regression is a common procedure for variable selection for multiple QTL analysis in composite interval mapping (Jansen1993; Zeng 1994) and multiple-interval mapping (Kao et al. 1999). The QTL effects are treated as fixed effects in step-wise regression. After a single locus model is fitted, the residuals are examined for the presence of a second QTL, and so on. A drawback of step-wise regression is that effects are included and removed from the model according to somewhat arbitrary statistical thresholds. Because many markers are tested in QTL mapping, the process necessarily entails relatively stringent significance thresholds for marker inclusion in the model. The result is that too few QTL are identified as significant in a mapping population, and the effects of those identified QTL are overestimated (Beavis, 1994; Beavis 1998; Schön et al., 2004; Xu 2003a). In addition, the QTL that exceed the chosen significance threshold often jointly only account for a limited proportion of the genetic variance. All these issues limit the scope and potential impact of MAS based on QTL analysis by step-wise regression.

More recently, Bayesian variable selection via the reversible jump Markov chain Monte Carlo (MCMC) algorithm was developed for QTL analysis (Satagopan et al. 1996; Heath

1997; Sillanpaa and Arjas 1998). Similarly, Yi et al. (2003) applied a stochastic search variable selection (SSVS, a Bayesian variable selection via Gibbs sampler) method to QTL analysis. One advantage of the Bayesian approach is that it considers the marginal posterior distributions of all parameters, including QTL position and effects given the data (Satagopan et al 1996).

New developments in shrinkage estimation seek to avoid variable selection by including all markers as predictors in the model and shrinking the estimated effects toward zero, rather than choosing a "best" set among them. By treating allelic effects as random, rather than as fixed effects, the shortage of degrees of freedom for estimating effects is avoided. Ridge regression (Hoerl and Kennard, 1970) is a classical example of shrinkage estimation. In ridge regression, the least squares effect estimators $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ are replaced by $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$ (Whittaker et al., 2000). A high value for the parameter λ causes a penalty for large β, thereby avoiding inflated estimates. Ridge regression is akin to the solution of the random component of Henderson's mixed-model equations for best linear unbiased prediction (BLUP) (Gianola, 2001), with $\lambda = \sigma_e^2/\sigma_\beta^2$, where σ_e^2 is the residual, and σ_β^2 is the estimator variance. This approach has strong affinities with estimation of β using random Bayesian models that assume a prior distribution for β: $\beta_j \sim N(0,\sigma_\beta^2)$. This is also close to an assumption of the infinitesimal genetic model for quantitative traits, i.e., many genes with small effects scattered across the genome.

A weakness of the ridge regression solution for including all markers is that all marker effects are equally penalized. To remove this constraint, hierarchical Bayesian random models that allowed for a different variance for each β_i ($\sigma_{\beta i}^2$) have been developed (Meuwissen et al. 2001; Xu 2003b; ter Braak et al 2005). These Bayesian approaches are more realistic than assuming equal variance for each locus by equal penalization on marker effects in the ridge regression method. Genomic selection for MAS proposed by Meuwissen et al. (2001) assumed population-wide LD is available with genome-wide dense SNP panels. Their genetic model assumption is that there are few genes with large effects and many genes

with small effects, which is likely more close to the true nature of most polygenic traits. One version of genomic selection, Bayes-B, has been shown, under the conditions simulated, to have comparable prediction accuracy for breeding values as progeny testing (Meuwissen et al., 2001).

Motivated by the random-model approach of Meuwissen et al (2001), Xu (2003b) proposed a hierarchical Bayesian shrinkage model for linkage QTL mapping using family-wise LD from a bi-parental cross. This model also allowed for a different variance for the regression effect of each marker, denoted β_i, and distributed as N(0, $\sigma^2_{\beta i}$). An interesting feature of Xu's (2003b) model is that it severely shrinks marker effects toward zero, more so as the effects become small. A consequence of this severe shrinkage is that the model reverses the usual bias of QTL effect estimates: under Xu's model, small effects are underestimated while little bias is present in the estimates of large effects (Wang et al., 2005). Important improvements to the model were proposed by ter Braak et al. (2005) to assure the proper posterior and better convergence of marker variance.

Since genomic selection and Bayesian shrinkage estimation estimate all marker effects, these approaches can account for small as well as large QTL effects. With continuously decreasing marker cost, MAS from these approaches would be better alternatives than traditional MAS for quantitative traits.

Association Analysis

Traditional MAS used QTL analysis that is usually carried out in a single mapping family derived from bi-parental crosses. However the linkage phase of the marker with the QTL within one family is not applicable across the population level, which results in the inability to extrapolate QTL effects from one breeding cross to another. Plant breeders usually generate many families of relatively small size. Marker-assisted selection methods must harmonize with plant breeding practice. Combining information from multiple families or crosses has been shown to be a powerful approach for QTL mapping (Rebaï and Goffinet, 1993; Muranty 1996; Xie et al. 1998; Xu, 1998; Rebaï and Goffinet, 2000; and Verhoeven et al., 2006; Blanc et al. 2006). QTL found from multiple elite inbred line crosses have several advantages over those from bi-parental populations. First, using multiple families of crosses

increases the statistical inference space and may permit detection of QTL which are undetectable in a single-line cross, where two parents could be fixed for the same allele at a particular QTL. The second benefit is that the effects of multiple QTL alleles can be estimated in different genetic backgrounds and thus the possible interaction between QTL and genetic background is detectable. In this context, the improvement of a line by the introgression of a QTL allele into a new genetic background is more predictable.

However, the above multiple-family mapping only used the within-family LD generated by bi-parental crosses and the marker-QTL association is still not generalizable to the whole breeding population. In a collection of germplasm, association mapping, which utilizes population-wide LD and allows much finer mapping than standard bi-parental cross approaches, is a more powerful approach (Thornsberry et al., 2001). The possibly large number of historical recombinations makes it possible to map causal loci more accurately through association than traditional linkage analyses (Flint-Garcia et al., 2003). If enough lines from a germplasm pool are sampled to develop the multiple families, as in regular plant breeding practice, population-wide LD could be established with dense markers, which would be particularly useful for association mapping or MAS in plant breeding programs. In such populations, short range LD between marker alleles and causal alleles arises not from experimental crossing but from historical drift and mutation events. Markers in population-wide LD are closely linked with QTL and thus the linkage phase and the estimated marker effects will be more consistent across the breeding population. Although long range LD will be created within each family, within-family LD will likely cancel each other at the population level if mating among the sampled lines is random. As an example of this idea, Yu et al. (2008) proposed nested association mapping (NAM) to dissect quantitative traits in maize. In the NAM design, diverse founders are selected and a large set of related mapping progenies are generated. The founders have complete sequence or dense markers and the complete marker genotypes of the progenies are inferred through linkage using the sparse markers in the progenies. With this design, Yu et al. (2008) successfully found a large fraction of the simulated QTL. Association analysis from this type of multiple-family population circumvents the need for constructing genetic mapping populations and instead utilizes existing breeding populations. Finally, the breeding lines tested are those of the

breeding programs themselves, making the QTL inferences and MAS immediately applicable.

To avoid spurious association with subpopulation, association analysis must consider population stratification, like the family structure mentioned above. Two competing models, mixed model analysis and genomic selection (Meuwissen, et al. 2001) exist for accounting population structure. In mixed-model analysis, marker alleles are considered fixed effects and population structure is accounted for by a random effect (Kennedy et al., 1992). Fitting the relevant random effect requires determining the kinship of individuals observed (Lynch and Walsh, 1998; Yu et al., 2006), which can be obtained either from molecular markers or pedigree records (Ritland, 2000). Simulation studies have confirmed mixed-model analysis to be useful in both cross- and self-pollinated crops (Arbelbide et al., 2006; Yu et al., 2005). The analysis has also been successful with real data in maize (Parisseaux and Bernardo, 2004) and wheat (Arbelbide and Bernardo, 2006; Breseghello and Sorrells, 2006). A weakness of these analyses is that they evaluate QTL effects one marker at a time. Surprisingly, genomic selection (Meuwissen et al 2001) as an association analysis, works admirably under a complex pedigree context without taking into account the variable kinships among individuals in the pedigrees. Habier et al (2007) demonstrated that dense markers could capture genetic relationships among genotyped individuals. In this way, the need for a separate random effect accounting for population structure, or variable relatedness, is removed. Another explanation is that with genome-wide dense marker and population-wide LD, no genomic regions with un-explained genetic effects arise.

Dissertation Organization

Significant developments as described above in Bayesian QTL analysis have been developed. Plant breeders now face the challenges of how to incorporate successful QTL analysis methods into breeding practice, for example, how to generate mapping families from candidate parents, how to predict the value of a cross, and how to carry out MAS in whole breeding program, etc. The overall objective of this dissertation is therefore to integrate QTL analysis into plant breeding using Bayesian statistics. Three different perspectives were

studied: identification of optimal designs to generate multiple families for QTL mapping, cross prediction using QTL analysis results, and genomic selection for cultivated barley.

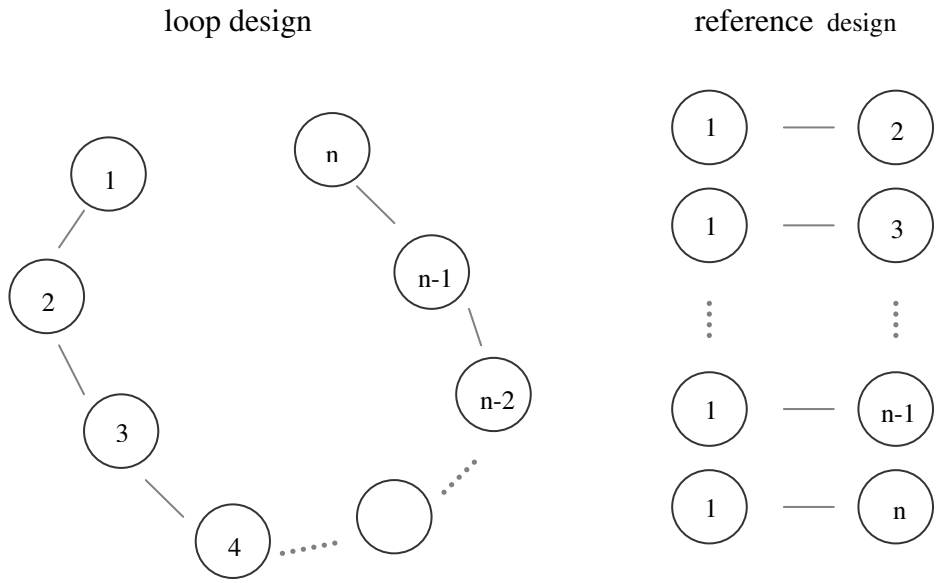

Figure 1. Two common mating designs in plant breeding. For the loop design, candidate parent 1 is cross to 2, 2 to 3, and so on, which results in n-1 families. For reference design, candidate parent 1 is cross to all other parents, which also generates n-1 families.

Each season, plant breeders usually generate many families each with small family size from candidate parents. It is important to investigate how different mating designs among these candidate parents affect QTL analysis. Loop and reference designs are two common mating designs (Figure 1) to generate multiple families in plant breeding. Chapter II explored within-family LD to study the impact of these two mating designs on QTL mapping in multiple families with a Bayesian variable selection approach. Assuming random parents from the germplasm were sampled, the impact of a loop mating design versus a reference mating one on power to detect QTL and to estimate QTL variance and position was studied.

In inbred line development, parents are crossed to generate segregating populations from which superior inbred progeny are selected. The usefulness of a particular cross does not depend on its mean progeny performance but on the performance of its best progeny (SCHNELL and UTZ, 1975), which was called the superior progeny value here. Plant breeders

tend to choose a cross that would have the higher probability of generating better superior inbred lines from a cross (Figure 2). In a typical breeding program, far too many crosses are possible between elite candidate parents for exhaustive evaluation. For example, among 50 elite parents there are 1225 possible crosses. Therefore it would be of great benefit if one could predict, among possible crosses, which ones are most likely to lead to superior inbred lines. Chapter III investigated cross prediction, using QTL analysis results from Bayesian shrinkage analysis that exploits within-family LD from a bi-parental cross. Theory was developed to predict the superior progeny value as a function of the mean of all progeny and of their standard deviation, using QTL analysis results. Different genetic conditions were used to assess the value of the genetic standard deviation in determining the usefulness of a cross.

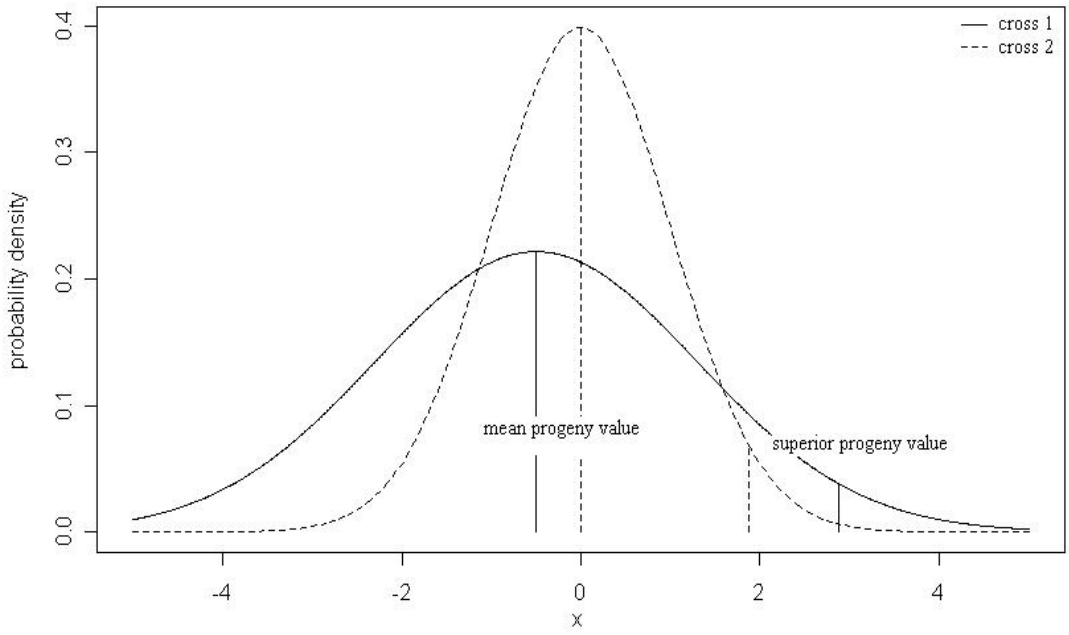

Figure 2. The usefulness of crosses. The graph shows the distribution of the progeny breeding values for two crosses. Although cross 2 has higher mean of all the progeny breeding values than cross 1, cross 1 has higher probability to obtain better superior progeny because cross 1 has larger genetic variation. Therefore cross 1 is preferred.

Chapter IV used barley SNP data to evaluate association-based genomic selection for spring two-row barley, assuming the existence of population-wide LD. We assessed the adequacy of current barley SNP density for genomic selection, and the impact of specific character in plant breeding relative to animal breeding upon genomic selection. The performance of three statistical analyses for genomic selection was compared: random regression best linear unbiased prediction (RR-BLUP) (Meuwissen et al 2001), Bayesian shrinkage estimation from ter Braak et al (2005) and Bayes-B (Meuwissen et al 2001).

Chapter V is general discussions followed by an appendix of the R code for genomic selection.

References

Allard, R.W. 1960 Principles of plant breeding. Wiley, New York.

Arbelbide, M., and R. Bernardo. 2006 Mixed-model QTL mapping for kernel hardness and dough strength in bread wheat. Theoretical and Applied Genetics 112:885-890.

Arbelbide, M., J. Yu, and R. Bernardo. 2006 Power of mixed-model QTL mapping from phenotypic, pedigree and marker data in self-pollinated crops. Theoretical and Applied Genetics 112:876-884.

Beavis, W. D., 1994 The power and deceit of QTL experiments: lessons from comparative QTL studies, p. 250-265, In D. B. Wilkinson, ed. Proceedings of the 49th Annual Corn and Sorghum Research Conference. American Seed Trade Association, Washington, D.C.

Beavis, W. D., 1998 QTL analyses: power, precision, and accuracy, pp. 145-162 in Molecular Dissection of Complex Traits, edited by A. H. PATERSON. CRC Press, New York.

Blair, M. W., A. J. Garris, A. S. Iyer, B. Chapman, S. K. Kresovich et al. 2003 High resolution genetic mapping and candidate gene identification at the xa5 locus for bacterial blight resistance in rice (Oryza sativa L.). Theor. And Appl. Genet. 107:62-73.

Blanc, G., A. Charcosset, B. Mangin, A. Gallais, and L. Moreau, 2006 Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. Theor. Appl. Genet. 113:206-224.

Brahm, L., T. Röcher, and W. Friedt 2000 PCR-Based Markers Facilitating Marker Assisted Selection in Sunflower for Resistance to Downy Mildew. Crop Sci. 40: 676-682.

Breseghello, F., and M.E. Sorrells. 2006 Association mapping of kernel size and quality in wheat (Triticum aestivum L.) cultivars. Genetics 172:1165-1177.

Chen, S., Lin XH, Xu CG, and Zhang Q 2000 Improvement of bacterial blight resistance of Minghui 63, an elite restorer line of hybrid rice, by molecular marker-assisted selection. Crop Science, 40, 239-244.

Dekkers, J. C. M. 2004 Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim. Sci. 82:E313-E328.

Erlich, H. A., and N. Amheim 1992. Genetic Analysis Using the Polymerase Chain Reaction. Annual Review of Genetics 26: 479-506.

FALCONER, D. S., and T. F. C. MACKAY, 1997 Introduction to Quantitative Genetics. Longman, New York.

Fehr, W.R., (ed.) 1984 Genetic contributions to yield gains of five major crop plants, pp. 1-101. Crop Science Society of America, Madison, USA.

Fernando, R.L., M. Grossman. 1989 Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. 21:467-477.

Flint-Garcia, S.A., J.M. Thornsberry, and E.S. Buckler, IV. 2003 Structure of linkage disequilibrium in plants. Annual Review of Plant Biology. Annual Review of Plant Biology. 54:357-374.

Gianola, D. 2001 Inferences about breeding values, p. 645-672, In D. Balding, et al., eds. Handbook of Statistical Genetics. John Wiley, New York.

Habier, D., R. L. Fernando and J. C. M. Dekkers. 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics. 177(4):2389-97.

Heath, S. C., 1997 Markov chain Monte Carlo segregation and linkage analysis of oligogenic models. Am. J. Hum. Genet. 61: 748-760.

Hoerl, A. E., and R. W. Kennard, 1970 Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12: 55-67.

Hospital, F. and A. Charcosset. 1997 Marker-assisted introgression of quantitative trait loci. Genetics 147:1469-1485.

Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics 135: 205-211.

Joshi, A.K., R. Chand, S. Kumar, and R.P. Singh 2004 Leaf tip necrosis: A phenotypic marker associated with resistance to spot blotch disease in wheat. Crop Sci. 44:792–796.

Kennedy, B.W., M. Quinton, and J.A.M. vanArendonk.1992 Estimation of effects of single genes on quantitative traits. Journal of Animal Science 70:2000-2012.

Kao C H, Z B Zeng and R D Teasdale 1999 Multiple interval mapping for quantitative trait loci. Genetics 152: 1203-1216.

Lande R, Thompson R 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743-756.

Lewontin, R. C and K. Kojima 1960 The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472.

Lynch M, and B Walsh 1998 Genetics and Analysis of Quantitative Traits. Sinauer Associates, Sunderland, MA.

Luo ZW, Hackett CA, Bradshaw JE, McNicol JW, Milbourne D, 2001 Construction of a genetic linkage map in tetraploid species using molecular markers. Genetics. 157:1369-1385.

Mason SL, Stevens MR, Jellen EN, Bonifacio A, Fairbanks DJ, McCarty RR, Rasmussen AG, Maughan PJ, 2005 Development and Use of Microsatellite Markers for Germplasm Characterization in Quinoa (Chenopodium quinoa Willd.) Crop Science 45:1618-1630.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Muranty H., 1996 Power of tests for quantitative trait loci detection using full-sib families in different schemes. Heredity 76: 156-165.

Parisseaux, B., and R. Bernardo, 2004 In silico mapping of quantitative trait loci in maize. Theoretical and Applied Genetics 109:508-514.

RebaÔ A, and Goffinet B, 1993 Power of tests for QTL detection using replicated progenies derived from a diallel cross. Theor. Appl. Genet. 86: 1014-1022.

Ritland, K. 2000 Marker-inferred relatedness as a tool for detecting heritability in nature. Molecular Ecology 9:1196-1204.

Satagopan, J. M., B. S. Yandell, M. A. Newton and T. G. Osborn, 1996 A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 144: 805-816.

Schon, C. C., H. F. Utz, S. Groh, B. Truberg, S. Openshaw, et al., 2004 Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 167:485-498.

Schnell, F.W., and H.F. Utz, 1975 F1-Leistung und Elternwahl Euphy-der Züchtung von Selbstbefruchtern. p. 243–248. Bericht über die Arbeitstagung der Vereinigung österreichischer Pflanzenzüchter. BAL Gumpenstein, Gumpenstein, Austria.

SERVIN, B., O. C. MARTIN, M. MEZARD, and F. HOSPITAL, 2004 Toward a theory of marker-assisted gene pyramiding. Genetics 168:513–523.

Sillanpaa M. J., and E. Arjas 1998 Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data†Genetics 148 (3): 1373-1388.

Singh S., Sidhu J. S., Huang N., Vikal Y., Li Z., Brar D. S., Dhaliwal H, Khush G. S. 2001 Pyramiding three bacterial blight resistance genes (xa-5, xa-13 and Xa21) using marker-assisted selection into indica rice cultivar PR106. Theor. Appl. Genet. 102:1011-1015.

Smith J. S. C., S. Kresovich, M.S. Hopkins, S.E. Mitchell, R.E. Dean, W.L. Woodman, M. Lee, and K. Porter, 2000 Genetic Diversity among Elite Sorghum Inbred Lines Assessed with Simple Sequence Repeats. Crop Sci. 40: 226-232.

ter Braak, C. J. F., M. P. Boer, and M. Bink, 2005 Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. Genetics 170:1435-1438.

Thornsberry, J. M., M. M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E. S. Buckler. 2001 Dwarf8 polymorphisms associate with variation in flowering time. Nature Genetics 28:286-289.

Verhoeven, K. J. F., J.-L. Jannink, and L. M. McIntyre, 2006 Using mating designs to uncover QTL and the genetic architecture of complex traits. Heredity 96:139-149.

Wang, G., T. S. Whittam, C. M. Berg, and D. E. Berg. 1993 RAPD (arbitrary primer) PCR is more sensitive than multilocus enzyme electrophoresis for distinguishing related bacterial strains. Nucleic Acids Res. 21:5930-5933.

Wang, H., Y. M. Zhang, X. M. Li, G. L. Masinde, S. Mohan, et al., 2005 Bayesian shrinkage estimation of quantitative trait loci parameters. Genetics 170:465-480.

Weeden, NF, and JF Wendel. 1990 "Genetics of plant isozymes". Pp. 46-72 in D. E. Soltis and P. S. Soltis, eds. Isozymes in plant biology. Chapman and Hall, London.

Whittaker, J. C., R. Thompson, and M. C. Denham, 2000 Marker-assisted selection using ridge regression. Genetical Research 75:249-252.

Willcox, M. C., M. M. Khairallah, D. Bergvinson, J. Crossa, J. A. Deutsch, et al 2002 Selection for resistance to southwestern corn borer using marker-assisted and conventional backcrossing. Crop Sci. 42: 1516-1528.

Winter P, and G. Kahl 1995 Molecular marker technologies for crop improvement. World J Microbiol Biotechnol 11:449-460.

Xie, C. Q., D. D. G. Gessler and S. Z. Xu, 1998 Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. Genetics 149: 1139-1146.

Xu, S., 1998 Mapping quantitative trait loci using multiple families of line crosses. Genetics 148: 517-524.

Xu, S., 2003a Theoretical Basis of the Beavis Effect. Genetics 165:2259-2268.

Xu, S., 2003b Estimating polygenic effects using markers of the entire genome. Genetics 163:789-801.

Yi, N., V. George and D. B. Allison, 2003 Stochastic search variable selection for identifying multiple quantitative trait loci. Genetics 164: 1129-1138.

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler, IV. 2006 A unified mixed-model method for association relatedness. Nat. Genet. 38:203-208.

Yu J, Holland J B, M D McMullen, and E S. Buckler 2008 Genetic Design and Statistical Power of Nested Association Mapping in Maize. Genetics 178: 539-551.

Zabeau, M and P. Vos. 1993 Selective restriction fragment amplification: a general method for DNA fingerprinting. European Patent Office, publication 0 534 858 A1, bulletin 93/13.

Zeng, Z-B., 1994 Precision mapping of quantitative traits loci. Genetics 136: 1457-1468.

CHAPTER II. COMPARISON OF TWO MATING DESIGNS FOR MULTIPLE-FAMILY QTL MAPPING

A paper to be submitted to *Crop Science*

Shengqiang Zhong and Jean-Luc Jannink

Abstract

Multiple-population quantitative trait locus (QTL) mapping integrates and uses information from many populations, and may improve QTL mapping and QTL-based breeding programs. Mating designs that maximize the value of multiple-population analysis have not been studied. Here, we simulated populations from two divergent designs, a reference design in which all parents are crossed to a reference parent, and a loop design in which parents are randomly ordered and crossed in a chain. Each design generated 12 families of 50 F_2 progeny. The genome consisted of seven 140 cM chromosomes, each carrying one QTL. The total QTL heritability was 0.42. The relative QTL detection power of the loop and reference designs depended on the detection threshold adopted. For typical thresholds (with an average detection power of 0.35) the loop design was most powerful, while for more stringent thresholds the reference design was most powerful. In all cases, the loop design gave more accurate estimates of QTL position and effect size. In general, we would recommend the loop design over the reference design for multiple-population QTL mapping.

Introduction

Characterizing the genetic architecture of a population requires detecting QTL and estimating the variance they generate. Traditional QTL mapping methods developed for progenies from a cross between two inbred parents have the problem of poor generalizability of the findings (Muranty, 1996; Xu, 1998). Comparisons of QTL mapping results from different populations suggest that, despite of some consensus QTL intervals across different mapping populations, a considerable amount of the QTL effects are family-specific (Welz and Geiger 2000; Kolb et al 2001; Kamoshita et al 2002; Clancy et al 2003; Chardon et al. 2004). Multiple-family QTL mapping, first proposed by Muranty (1996), can avoid this problem. Studies have shown that using multiple line crosses broadens the parameter inference space to the reference population, leads to more generalizable inferences about QTL, and can increase QTL detection power (Muranty, 1996; Xu, 1998; Rebaï and Goffinet, 2000; Jannink, 2001).

Important theoretical developments have been made on QTL detection methods in interconnected families. Statistical methods can be distinguished into three categories: regression analyses (Rebaï and Goffinet 2000), maximum-likelihood methods (Liu and Zeng, 2000), and Bayesian approaches (Xu, 1998; Jannink and Wu, 2003). Bayesian methods have been demonstrated to be able to accommodate more complicated models, such as variable QTL number (Satagopan, 1996; Sillanpaa, 1998; Xu, 1998) and variable allele configuration models (Jannink and Wu, 2003). In these analyses, QTL allele effects have been considered fixed (Rebaï and Goffinet, 1993; Rebaï et al., 1994; 2000; Liu and Zeng, 2000) or random (Xu, 1998; Xie et al., 1998; Jannink and Wu, 2003). Fixed allele effect models relax the assumption of normally distributed allele effects but face the problem of estimating a large number of parameters when many families are analyzed jointly and all parental alleles are assumed to have distinct effects. Random allele effect models estimate only a single allelic effect variance, so the number of parameters per QTL is independent of the number of families.

Given the growing recognition of the importance of multiple-family QTL analysis (Blanc et al 2006), evaluation of appropriate experimental mating designs to optimize the estimation

of relevant parameters are needed. Researchers have investigated how different sampling strategies (family number vs. family size) might affect QTL detection power (Xie et al 1998; Xu 1998; Rao and Li 2000) and QTL allelic number and variance estimation (Wu and Jannink 2004). These studies showed that the power of QTL detection was higher with an intermediate number of families than with a small number of large families or a large number of small families. As to the effect of mating design on QTL mapping, Muranty (1996) showed that when the number of families and the number of offspring per family were held constant, QTL detection power was affected by the number of outbred parents used in the design but the arrangement of the outbred parents in specific diallel designs had little impact. Wu and Jannink (2004) investigated a circulant diallel mating design. This design is sometimes used in plant breeding practice (e.g., DeKoeyer and Stuthman, 1998) and would therefore be a logical choice for combining QTL analyses with further advancing a breeding program. For the specific purpose of distinguishing among QTL alleles carried by different parents, however, they pointed out that the circulant diallel design may not be the most effective crossing scheme. Alleles carried by two parents will be contrasted with the greatest power if those two parents are crossed directly. This reasoning suggests that all pair-wise crosses be performed among parents sampled from the reference population, leading to a half-diallel design. Verhoeven et al. (2006) explored QTL detection power with a fixed number of parents under different mating designs. For a fixed experiment size, they found that QTL detection power was greatest for the mating design with the fewest but largest families (circulant diallel mating design) while power was lowest for the design with the most, smallest families (the half diallel). This study showed that half-diallel design analyzing a high number of families was not the best. Another alternative is to cross all sampled parents to the same reference parent. This approach does not allow the direct contrasts among QTL alleles that the half-diallel allows, but instead an indirect contrast relative to a very-well characterized reference allele.

In this paper, we use a Bayesian QTL mapping method to contrast two mating designs, the circulant diallel design, or loop design, and the reference design. The criteria used to assess mating design performance were QTL detection power, accuracy of QTL position estimation, and accuracy of allelic effect variance estimation. The analysis presented here

assumes that a fixed number of individuals have been randomly picked from a base population and inbred to homozygosity to be used as parents in the mating designs. These homozygous parents are then crossed to generate F_2 QTL mapping families of a fixed size.

Methods

Simulations

(i) Mapping families: The mapping families were generated using either a modified loop design or a reference design. In the loop design, inbred founders were randomly ordered and each founder was mated with its immediate neighbors in the list. The first and the last parents on the list were mated once, while all other parents were mated twice. In the reference design, a single inbred founder was chosen at random to be the reference parent, and all other founders were crossed to it. Both designs generate $P - 1$ families from P founding inbred parents. Here, 13 inbred founders were used to generate 12 F_2 mapping families, with 50 F_2 progenies in each family leading to a fixed experimental size of 600.

(ii) Genome, markers and QTL: The simulated genome consisted of 7 chromosomes of a length of 140 cM for a total genome size similar to barley. The marker spacing was 10 cM, markers were co-dominant and informative in all crosses. Seven additive effect QTL were simulated on the genome with one QTL on each chromosome. As in Verhoeven et al. (2006), each inbred founder was simulated to carry a unique QTL allele. The effects of these QTL in terms of the additive variance contributed to the trait varied in a geometric series (Lande and Thompson, 1990). The percent of the phenotypic variance generated by the QTL in the random-mating F_2 population were 12, 9, 7, 5, 4, 3, and 2 for a total heritability of 0.42. All QTL were simulated at 45 cM from the end of their chromosome.

QTL model and statistical analysis

(i) *QTL model* Consider a simple additive model for F_2 populations, where the trait of interest is affected by J QTL. Let N be the total number of F_2 progeny that were produced by K F_1 individuals, and the latter were derived from P inbred founders. Assume that each founder carries a distinct allele at each QTL. The vector of observed phenotypes y is modeled as

$$y = X\beta + \sum_{j=1}^{J} Q_j \alpha_j + \varepsilon \quad (1)$$

where X is an $N \times K$ design matrix relating an F_2 progeny to the F_1 family from which it is derived, β is a $K \times 1$ vector of family means, Q_j is a $N \times P$ matrix relating a progeny to the allele that QTL j carries, α_j is a $P \times 1$ vector of allelic effects at QTL j, and $\varepsilon \sim N(0, I\sigma^2)$ is a $N \times 1$ vector of residuals.

(ii) Statistical *model* We observe trait values (y), family structure (X) and marker genotypes (M). Marker genotypes are considered fixed and the analysis is conditioned on the marker genotypes observed. We wish to infer the number of QTL (J), QTL positions (Λ_j), QTL genotypes of each progeny (Q_j), allelic effect variance for each QTL (σ_j^2), allelic effects (α_j), family means (β), and residual variance (σ^2). We employ a Bayesian analysis with the following priors. The prior for J is Poisson with mean λ, the expected QTL number in the genome. The prior for Λ_j is uniform over the genome so that $p(\Lambda_j)$ = (genome size)$^{-1}$. The prior for Q_j follows the rules of Mendelian segregation and recombination conditioned on Λ_j and on flanking markers. Allelic effects are treated as random, where the prior for α_j is normal with mean zero and variance σ_j^2. σ_j^2 is considered a random variable to be estimated with uniform prior distribution $\sigma_j^2 \sim unif(0, \sigma_{\max}^2)$, where $\sigma_{\max}^2$ is a maximum value that we believe allelic effect variance can take.

Denote $\theta = (\beta, \sigma^2, \{\Lambda_j\}_{j=1}^J, \{Q_j\}_{j=1}^J, \{\alpha_j\}_{j=1}^J, \{\sigma_j^2\}_{j=1}^J)$ the vector of all unobservable parameters, excluding the number of QTL. The joint posterior density of all unobservable parameters (J, θ) given the observable (y, X, M) and prior information is

$$p(J, \theta \mid y, X) \propto p(y \mid \theta, J)\, p(J)\, p(\beta)\, p(\sigma^2)\, p(\Lambda) \prod_{j=1}^{J} p(Q_j \mid \Lambda, M) \prod_{j=1}^{J} p(\alpha_j \mid \sigma_j^2) p(\sigma_j^2) \quad (2)$$

where $p(y \mid J, \theta)$ represents the likelihood assuming $\varepsilon \sim N(0, I\sigma^2)$ and $p(*)$ is the prior distribution for parameter *, with $p(Q_j \mid \Lambda, M)$ being the prior distribution for genotypes at QTL j conditional on QTL positions and flanking markers genotypes, $p(\alpha_j \mid \sigma_j^2)$being the normal prior distribution for allele effects conditional on allelic effect variance, and $p(J)$ being the Poisson prior with mean parameter λ.

(iii) Model *implementation by Markov chain Monte Carlo (MCMC).* The analysis was accomplished by repeatedly sampling from the posterior distribution using Markov chain Monte Carlo techniques. Parameter estimates were then provided by their marginal posterior distributions. We employed a scalar Metropolis-Hastings algorithm where each parameter in θ was sampled in turn, considering all other parameters fixed (Gilks et al., 1996). The exception to this scalar implementation was that when updating QTL position, QTL genotypes, allelic variance and allelic effects were sampled jointly with the position. Note that the dimension of the parameter vector θ changes when the number of QTL J is changed. We thus used a reversible jump Metropolis-Hastings step to move between different numbers of QTL by either adding a new QTL to the model or dropping an existing QTL from the model (Jannink and Fernando, 2004).

After all parameters were initialized from their priors, a complete MCMC iteration consisted of the following steps:

- Update QTL allele effects α_j for each QTL j and each inbred founder P
- Update QTL position: update position Λ_j and genotypes Q_j, allelic effect variance σ_j^2, and allele effects α_j jointly for each QTL j
- Update family means β, and residual variance σ^2
- Update the number of QTL J

For a given number of QTL, the steps for updating QTL allele effects, position, family means, and residual variance were described in detail in Wu and Jannink (2004). The methods of updating the number of QTL under a random allele effect model were adapted from Jannink and Fernando (2004).

Two hundred simulations were run for both mating designs under two different models of analysis: either by fixing the correct QTL number, one per chromosome (Fixed Model, FM), or by allowing the number of QTL to be variable (Variable Model, VM) by reversible jump MCMC. For each analysis, four chains were run with 15,000 and 30,000 burn-in iterations for the FM and VM models, respectively, to ensure chain convergence. Subsequently, with a thinning value of 1, 30,000 samples were collected for each chain after burn-in, for a total of 120,000 samples per analysis.

QTL detection power, allelic variance and position estimation

Location-wise QTL intensity was defined in 1 cM bins as the probability that a QTL was sampled in that bin over the 120,000 samples. Sliding windows of 10 cM over which intensity was integrated were used to evaluate QTL detection power, accuracy of map position estimation, and location-wise QTL variance. Integrated intensity and average QTL variance centered at two kinds of positions were obtained. One was the true QTL position (pos_{True}). The other was obtained by identifying the 10 cM window where integrated intensity was maximized ($pos_{\max}$). The summary statistic $pos_{\max}$ was also used as an estimator of QTL position. An additional estimator of QTL position was the expected posterior QTL position (*Epos*) for each chromosome, as follows. Assuming chromosome length chrL, $Epos = \frac{\sum_{i=1}^{chrL} i * intensity_i}{\sum_{i=1}^{chrL} intensity_i}$, where $intensity_i$ was the location-wise intensity at bin *i*.

We determined the statistical power for detecting a QTL by counting the number of runs where the 10 cM integrated intensity surrounding the true position was greater than a certain threshold, or by counting the number of runs where the integrated intensity surrounding $pos_{\max}$ was greater than a certain threshold and $pos_{\max}$ was within 10 cM of the true position. Let π_0 be the prior QTL intensity. In the present analysis, $\pi_0 = \bar{J}$ [QTL]/9.8 [Morgans], where $\bar{J}$ is the average number of QTL (parameter *J* above) over MCMC iterations, and 9.8 is the simulated genome size in Morgans. For the FM model, $\bar{J}$=7 and π_0=7 [QTL]/9.8 [Morgans]=0.714 [QTL/Morgan]; for the VM model, $\bar{J}$ needed to be calculated for each analysis. Using π_0, we defined a detection threshold unit, T = π_0x0.10, where 0.10 was the length in Morgans of the interval over which QTL intensity was integrated. Under the FM model, T = 0.0714. Arbitrarily, we choose different thresholds at which to declare a QTL present, from 1T to 15T. In the case of 1T, the analysis was minimally stringent, in that it sufficed for the QTL intensity surrounding the true (simulated) QTL position to be greater than the prior intensity for a QTL to be declared present.

To quantify the accuracy of QTL allelic variance and position estimation, the mean squared error (MSE), $MSE(\hat{\theta}) = Bias^2(\hat{\theta}) + Var(\hat{\theta})$, the sum of the squared measurement bias and the measurement variance were used. Measurement variance was the usual variance among replicate estimates. Bias was calculated as $Bias(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta})_r - \theta$ where R was the number of replicate simulations, $\hat{\theta}_r$ was the estimate in simulation r, and θ was the true value. Two kinds of QTL allelic variance estimator was obtained as the average QTL variance at pos_{True} and $pos_{\max}$. Two kinds of estimators, $pos_{\max}$ and *Epos*, were utilized to estimate QTL position..

Results

QTL number in reversible jump MCMC

The average QTL numbers of VM model were 5.30 ± 0.07 and 5.86 ± 0.07 in the reference and loop design, respectively. These estimates were smaller than the true QTL number (7), possibly due to the small allelic variance of some QTL in the genome. Nevertheless, the VM model selected a number of QTL closer to the true number when analyzing the loop design than when analyzing the reference design, indicating that the loop design was better than the reference design from this perspective.

QTL detection power

Using a 3T threshold, detection power from pos_{True} was 4% and 6% higher for the loop design than for the reference design, under the FM and VM models, respectively (Figure 1a); detection power from $pos_{\max}$ was similar for two mating designs under the FM model but 5% higher for the loop design under the VM model (Figure 1b). Interestingly, as the threshold that was used to declare detection increased and the detection power decreased, the reference design increased in power relative to the loop design, such that for high thresholds (> 7*T*), the reference design was more powerful than the loop design for both the FM and the VM model (Figure 1a and 1b).

QTL allelic variance

The loop design had lower mean squared error for QTL allelic variance than the reference

design (Figure 2a and 2b). The MSE from pos_{True}, was 8% lower for the loop than the reference design under both FM and VM. The MSE from $pos_{\max}$, was 17% lower for the loop design than for the reference design under FM, and 23% under VM.

QTL position

The MSE for QTL position was smaller for the loop than the reference design for both FM and VM and for both position estimates, *Epos* and $pos_{\max}$ (Figure 3a and b). The MSE using *Epos* was 5% lower for the loop design than for the reference design under FM, and 9% under VM. The MSE using $pos_{\max}$ was 6% lower for the loop design than for the reference one under FM, and 11% under VM.

Discussion

Joint analysis of multiple families, first introduced by Muranty (1996), extends QTL mapping to a broader inference space with respect to the reference population and can reduce the problem of non-segregating QTL in single cross (Xu, 1998). Thus multiple-family QTL mapping can improve QTL detection at the population level and be incorporated into practical breeding programs for individual genotypic value prediction (Verhoeven et al, 2006).

For a given experimental size, the question of how many parents should be sampled and how the mating designs should be arranged becomes important in QTL analysis. Wu and Jannink (2004) studied how the number of parents sampled affected the estimation of QTL allele number at a given QTL. Verhoeven et al (2006) explored how the number of families in diallel designs influenced QTL detection, given a fixed number of parents and equal contributions of each parent to the number of the families. They found that QTL detection power was greatest for the mating design with the fewest but largest families (loop design). Here we used fixed experimental size, fixed parent number, but compared how unequal contributions of the parents to the number of the families might impact the QTL detection. Two extreme situations were used: one was the loop design, where each parent (except the first and last one of the random ordered parents) equally contributes to two of the families; another was the reference design, where one parent contributes to all of the families and the

rest to just one family. It is interested to investigate other intermediate situations.

Although the reference design offers a well-characterized reference allele with which to contrast the other alleles and might therefore increase QTL detection efficiency (Wu and Jannink 2004), the results presented here suggest that it causes insufficient representation of the other alleles, which naturally generates larger variance for parameter estimation from experiment design view. This reduced representation appears to reduce the efficiency of the analysis from several perspectives. In particular, simulations showed that the loop design estimated QTL position and QTL variance with lower mean squared error than the reference design. Both designs detected QTL with similar power over different thresholds, but the loop design showed higher power when lower detection thresholds were adopted while the reference design showed higher power when higher detection thresholds were adopted.

To understand this phenomenon, we analyzed the mean and coefficient of variation (CV) of the 10 cM integrated intensity either from pos_{True} or from $pos_{\max}$. The mean integrated intensity of QTL was similar for the loop and reference designs (Figure 4a and b). Mixed model analysis showed that design had no overall effect on integrated intensity from pos_{True} ($P = 0.421$), but integrated intensity from $pos_{\max}$ in the reference was 4% higher than in the loop design ($P = 0.046$). The CV of the integrated intensity was consistently larger in the reference design than in the loop design (Figure 5a and b). Analysis of the residual variances from the mixed models (Littell and Ramon, 1996) showed that the residual variance in the reference design was larger than that in the loop design for integrated intensity from both pos_{True} ($P < 0.001$) and $pos_{\max}$ ($P < 0.001$).

We hypothesize that this higher CV arose for the following reason. If the parent randomly selected to be the reference has extreme high or low allelic effects, that parent will offer a high contrast in QTL mapping, leading to high QTL intensities. On the other hand, if the parent chosen to be the reference is intermediate, it will offer a poor contrast to the other parents, leading to low intensities at the true QTL positions. Randomly selecting the parent, as occurred here, therefore leads to high variance in intensity. When the detection threshold chosen is high, this high variance can lead to greater detection power because some simulations will deviate from the mean intensity also toward the high end. Therefore for

reference design, it would be preferred to select a reference parent with high or low breeding value for single trait QTL mapping, although it might not be easy to select such a reference parent for QTL mapping in multiple traits.

In conclusion, because the loop design more uniformly represents alleles sampled from a base population it offers advantages over the reference design in terms of characterizing QTL. In the present study, these advantages were small but detectable: the loop design gave more accurate estimates of the QTL position and of the variance generated by the QTL. For QTL detection power, the results were somewhat more equivocal. Nevertheless, we would recommend the loop design over the reference design because low CV of QTL intensity, as given by the loop design, will result in more consistent QTL detection performance.

References

Blanc G, Charcosset A, Mangin B, Gallais A, Moreau L (2006) Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize Theor Appl Genet 113:206–224.

Chardon F, Virlon B, Moreau L, Falque M, Joets J, Decousset L, et al. (2003). Comparative mapping of beta-amylase activity QTLs among three barley crosses. Crop. Sci. 43: 1043-1052.

De Koeyer, DL, and D.D. Stuthman (1998). Continued response through seven cycles of recurrent selection for grain yield in oat (Avena sativa L.). Euphytica 104:67-72.

Gilks WR, Richardson S, Spiegelhalter DJ (1996) Introducing Markov chain Monte Carlo. In: Gilks WR, Richardson S, Spiegelhalter, DJ (eds) Markov chain Monte Carlo in practice. Chapman and Hall, London, pp 1–19.

Jannink JL, Fernando RL (2004). On the Metropolis-Hastings acceptance probability to add or drop a quantitative trait locus in Markov chain Monte Carlo-based Bayesian analyses. Genetics 166: 641–643.

Jannink JL, Jansen R (2001). Mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics 157: 445-454.

Jannink JL, Wu XL (2003). Estimating allelic number and identity in state of QTLs in interconnected families. Genet Res 81: 133–144.

Kamoshita A, Wade LJ, Ali ML, Pathan MS, Zhang J, Sarkarung S et al (2002). Mapping QTLs for root morphology of a rice population adapted to rainfed lowland conditions. Theor. Appl. Genet. 104: 880-893.

Kolb FL, Bai GH, Muehlbauer GJ, Anderson JA, Smith KP, Fedak G (2001). Host plant resistance genes for fusarium head blight: Mapping and manipulation with molecular markers. Crop. Sci. 41: 611-619.

Lande R, Thompson R (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743-756.

Littell, Ramon C. SAS system for mixed models. Cary, N.C. : SAS Institute, Inc., c 1996.

Liu YF, Zeng ZB (2000). A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines. Genet. Res. 75: 345-355.

Muranty H (1996). Power of tests for quantitative trait loci detection using full-sib families in different schemes. Heredity 76: 156-165.

Rao SQ, Li X (2000). Strategies for genetic mapping of categorical traits. Genetica 109: 183-197.

Rebaï A, Goffinet B (2000). More about quantitative trait locus mapping with diallel designs. Genet. Res. 75: 243-247.

Rebaï A, Goffinet B, Mangin B, Perret D (1994) Detecting QTLs with diallel schemes. In van Ooijen JW, Jansen J (eds) Biometrics in Plant Breeding: Applications of Molecular Markers, 9th meeting of the EUCARPIA, Wageningen, the Netherlands, CPRO-DLO, pp. 170-177.

Satagopan JM, Yandell YS, Newton MA, Osborn TC (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 144: 805-816.

Sillanpaa MJ, Arjas E (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics 148:1373-1388.

Verhoeven, KJF, Jannink JL, McIntyre, LM. (2006). Using mating designs to uncover QTL and the genetic architecture of complex traits. Heredity. 96:139-149.

Welz HG, Geiger HH (2000). Genes for resistance to northern corn leaf blight in diverse maize populations. Plant Breeding 119: 1-14.

Wu XL, Jannink JL (2004). Optimal sampling of a population to determine QTL location, variance, and allelic number. Theor. Appl. Genet. 108: 1434-1442.

Xie C, Gessler DD, Xu S (1998) Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. Genetics 149:1139-1146.

Xu S (1998). Mapping quantitative trait loci using multiple families of line crosses. Genetics 148: 517-524.

Figures

Figure 1. QTL detection power under different thresholds for the reference (Ref) and Loop designs for models with fixed (FM) or variable (VM) numbers of QTL fitted. The results were from 200 replicate simulations and the error bars are standard error. Power was estimated using 10 cM integrated intensity averaged over the seven simulated QTL at their a. true position , and b. estimated QTL position.

Figure 2. Mean square error (MSE), squared bias and measurement variance (VAR) of posterior QTL allelic variance for the reference (Ref) and Loop designs for models with fixed (FM) or variable (VM) numbers of QTL fitted. The above three statistics (MSE, Squared Bias and Var) were averaged over the seven simulated QTL at their a. true position, and b. estimated QTL position. Each standard error bar in the graph represents an average of each statistic over all seven QTL and all replicate simulations.

Figure 3. Mean square error, squared bias and measurement variance of QTL position estimation for the reference (Ref) and Loop designs for models with fixed (FM) or variable (VM) numbers of QTL fitted. The above three statistics (MSE, Squared Bias and Var) were averaged over the seven simulated QTL at their a. estimated expected QTL position $Epos$, and b. another estimated QTL position $pos_{\max}$. Each standard error bar in the graph represents an average of each statistic over all seven QTL and all replicate simulations.

Figure 4. Average 10 cM integrated intensity at the a. expected QTL position $Epos$, and b. another estimated QTL position $pos_{\max}$ across 200 replicate simulations for the reference (Ref) and Loop designs for models with fixed (FM) or variable (VM) numbers of QTL fitted. QTL allelic variance simulated became smaller from QTL 1 to QTL.

Figure 5. Coefficient of variation of 10 cM integrated intensity at their a. estimated expected QTL position $Epos$, and b. another estimated QTL position $pos_{\max}$ across 200 replicate simulations for the reference (Ref) and Loop designs for models with fixed (FM) or variable (VM) numbers of QTL fitted. QTL allelic variance simulated became smaller from QTL 1 to QTL 7.

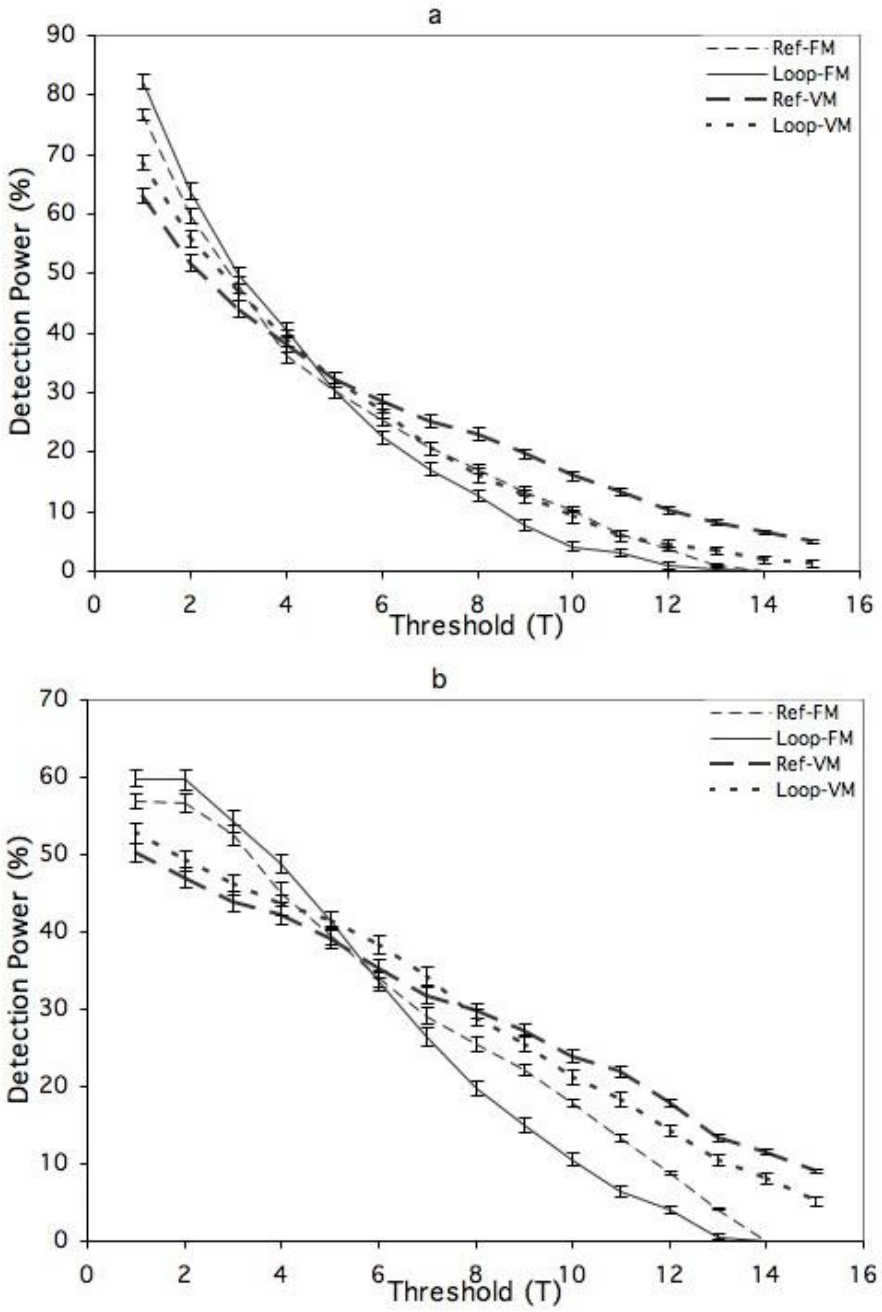
a
Ref-FM
Loop-FM
Ref-VM
Loop-VM
Detection Power (%)
Threshold (T)
b
Ref-FM
Loop-FM
Ref-VM
Loop-VM
Detection Power (%)
Threshold (T)

Figure 1

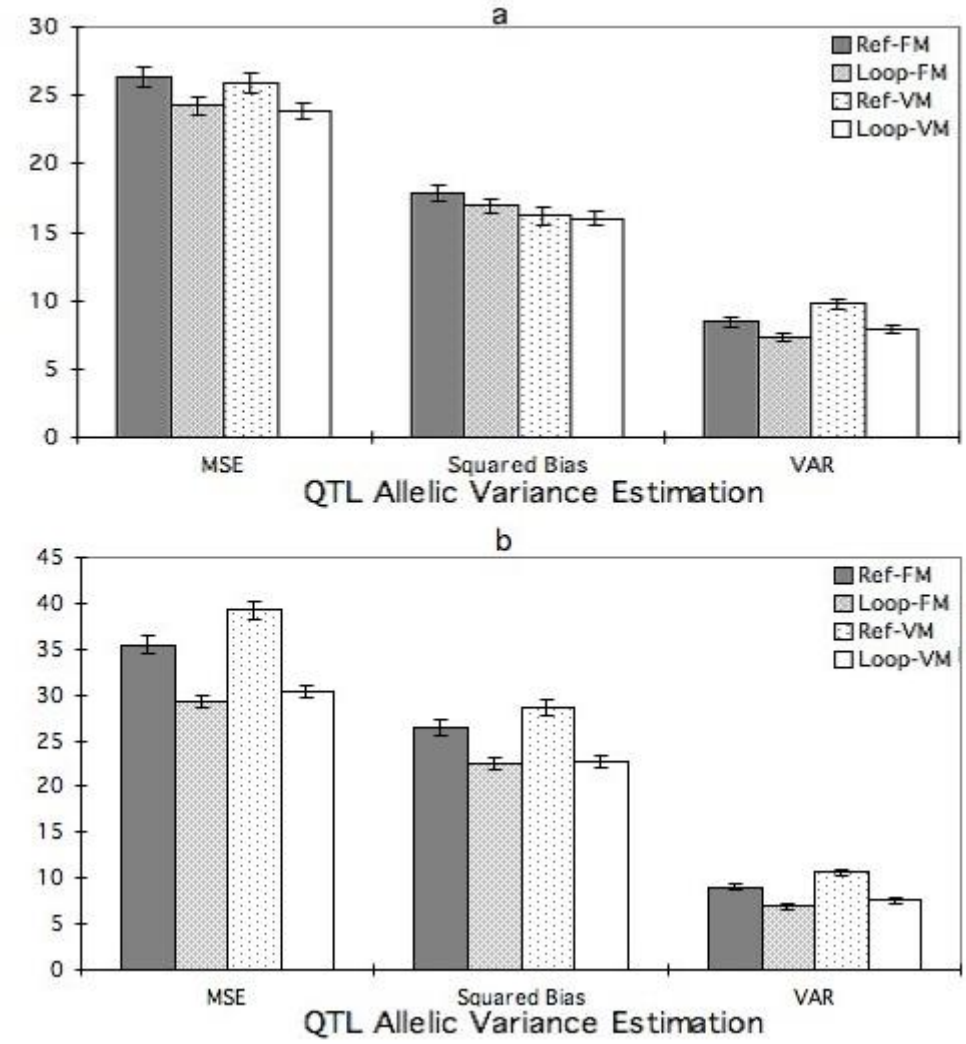
a
30
25
20
15
10
5
0
Ref-FM
Loop-FM
Ref-VM
Loop-VM
MSE
Squared Bias
VAR
QTL Allelic Variance Estimation
b
45
40
35
30
25
20
15
10
5
0
Ref-FM
Loop-FM
Ref-VM
Loop-VM
MSE
Squared Bias
VAR
QTL Allelic Variance Estimation

Figure 2

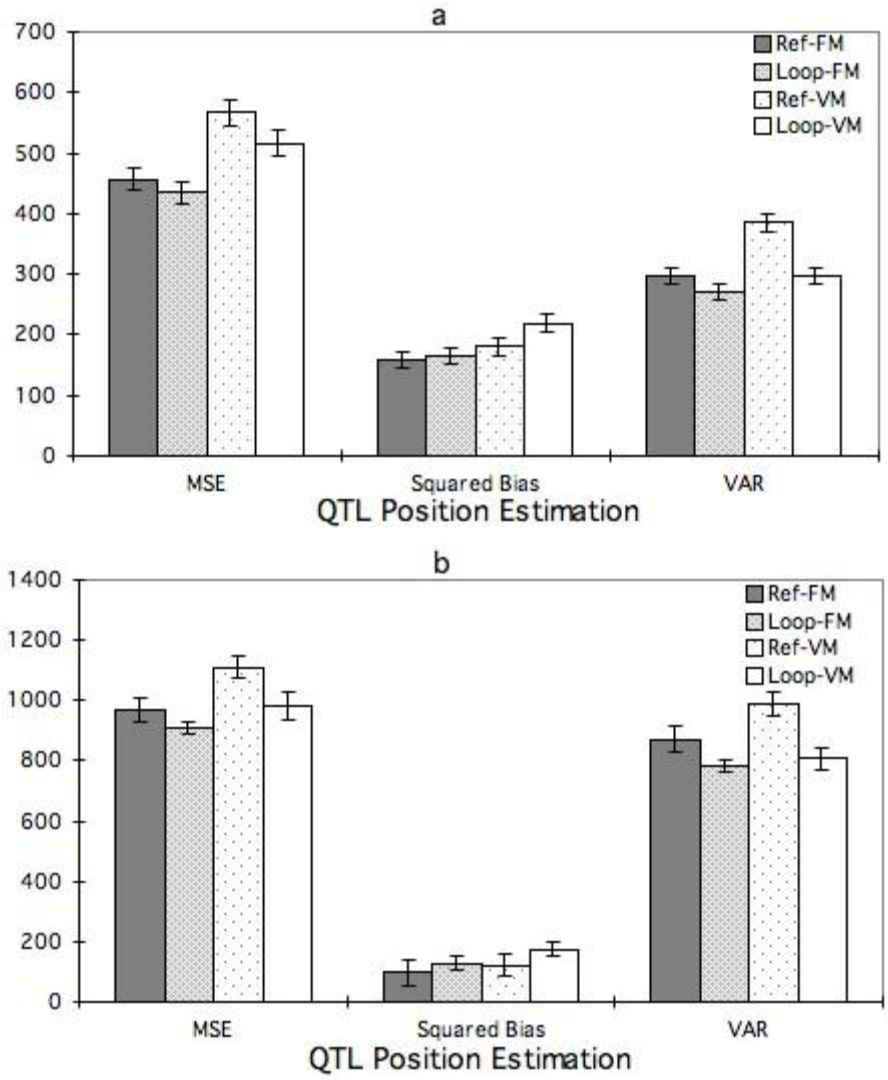
a
700
600
500
400
300
200
100
0
Ref-FM
Loop-FM
Ref-VM
Loop-VM
MSE
Squared Bias
VAR
QTL Position Estimation
b
1400
1200
1000
800
600
400
200
0
Ref-FM
Loop-FM
Ref-VM
Loop-VM
MSE
Squared Bias
VAR
QTL Position Estimation

Figure 3

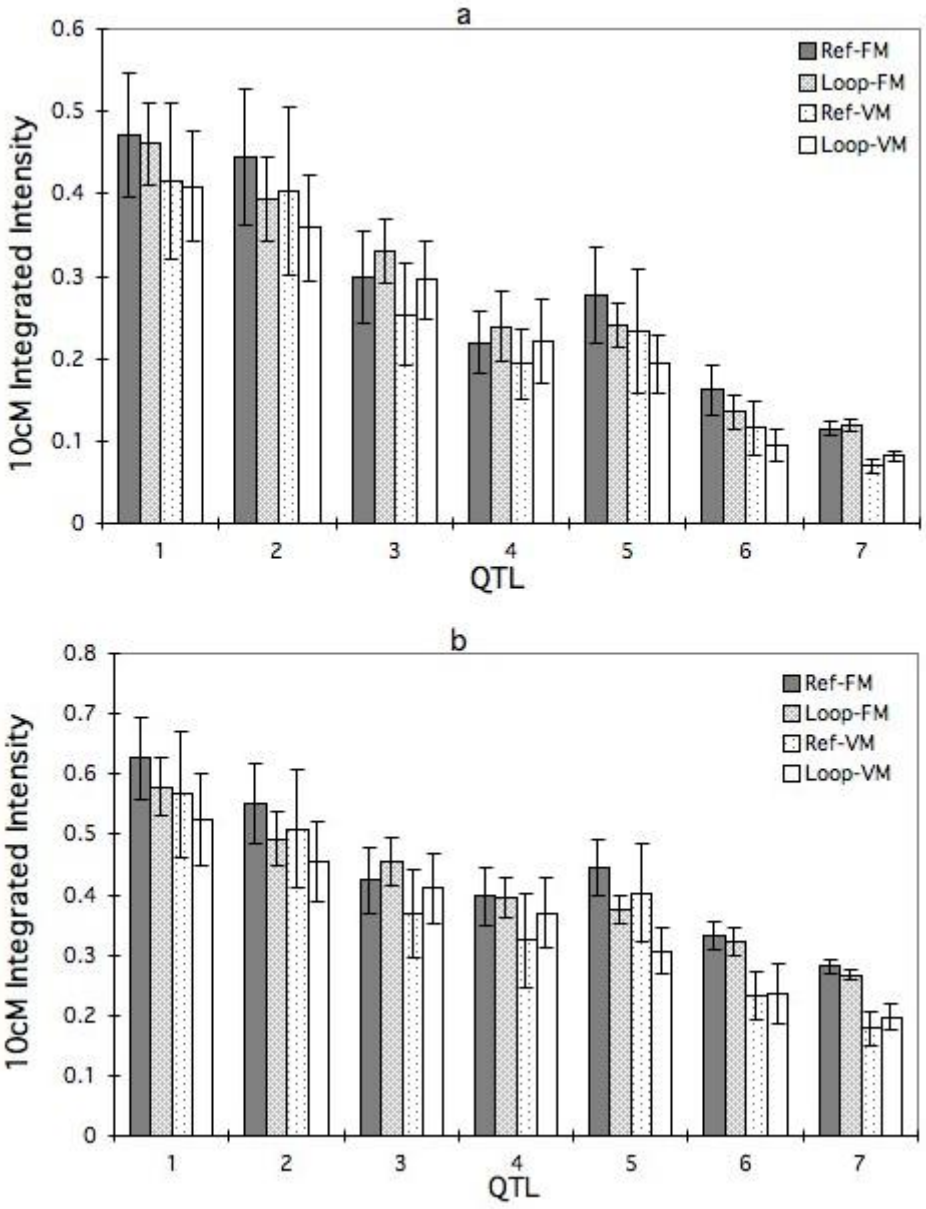
a
10cM Integrated Intensity
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
QTL
Ref-FM
Loop-FM
Ref-VM
Loop-VM
b
10cM Integrated Intensity
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
QTL
Ref-FM
Loop-FM
Ref-VM
Loop-VM

Figure 4

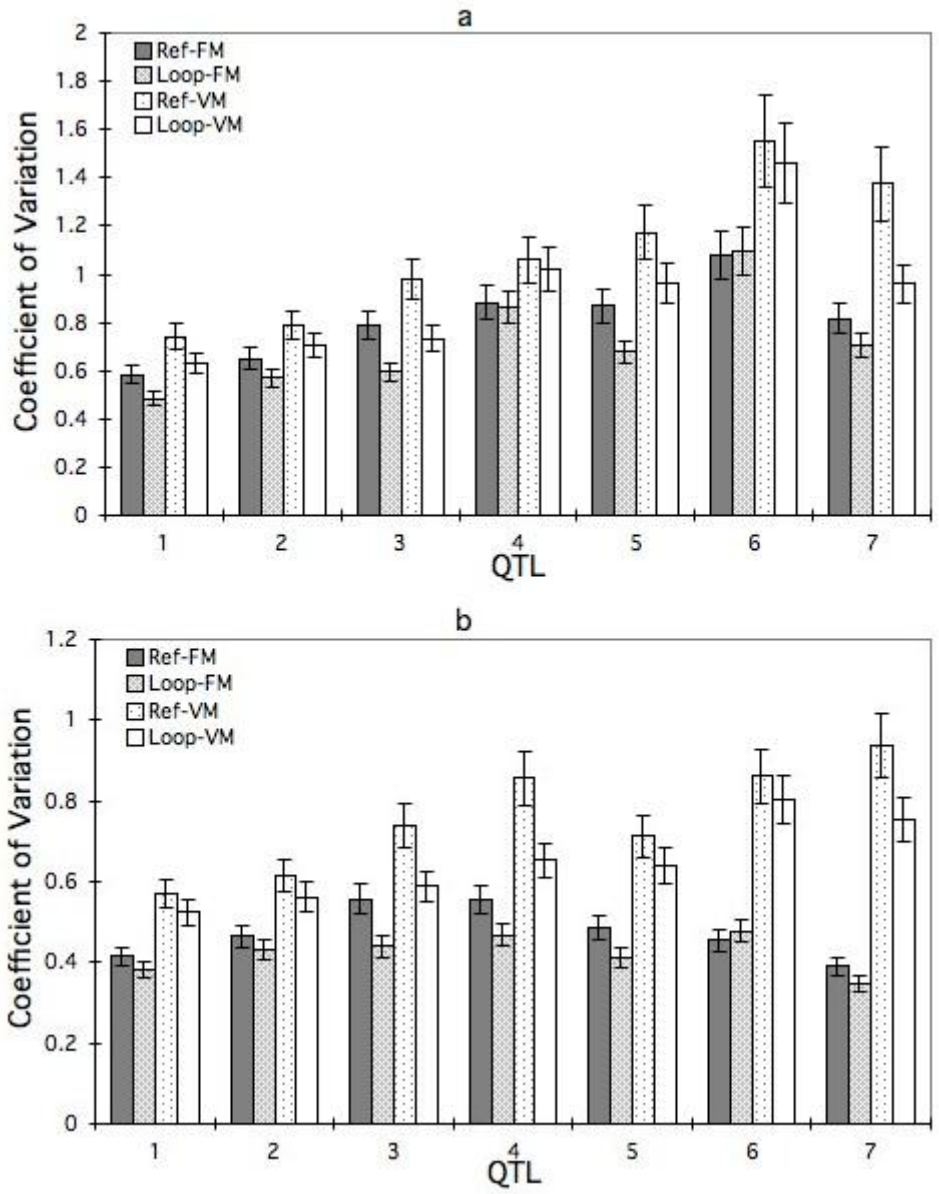
a
Ref-FM
Loop-FM
Ref-VM
Loop-VM
Coefficient of Variation
QTL
b
Ref-FM
Loop-FM
Ref-VM
Loop-VM
Coefficient of Variation
QTL

Figure 5

CHAPTER III. USING QTL RESULTS TO DISCRIMINATE AMONG CROSSES BASED ON THEIR PROGENY MEAN AND VARIANCE

Published in *Genetics*

Shengqiang Zhong[*] and Jean-Luc Jannink[†1]

Abstract

In order to develop inbred lines, parents are crossed to generate segregating populations from which superior inbred progeny are selected. The value of a particular cross thus depends on the expected performance of its best progeny, which we call the superior progeny value. Superior progeny value is a linear combination of the mean of the cross's progeny and their standard deviation. In this study we specify theory to predict a cross's progeny standard deviation from QTL results and explore analytically and by simulation the variance of that standard deviation under different genetic models. We then study the impact of different QTL analysis methods on the prediction accuracy of a cross's superior progeny value. We show that including all markers, rather than only markers with significant effects, improves the prediction. Methods that account for the uncertainty of the QTL analysis by integrating over the posterior distributions of effect estimates also produce better predictions than methods that retain only point estimates from the QTL analysis. The utility of including estimates of a cross's among-progeny standard deviation in the prediction increases with increasing heritability and marker density but decreasing genome size and QTL number. This

---

[*] Department of Agronomy, Iowa State University, Ames, Iowa 50011-1010, [†]USDA–ARS, U.S. Plant, Soil, and Nutrition Laboratory, Ithaca, New York 14583

Running head: Selection among crosses

Keywords: Genome-wide selection, Marker assisted selection, Progeny variance, Prediction accuracy, QTL analysis

[1] Corresponding author: Jean-Luc Jannink, USDA–ARS, U.S. Plant, Soil, and Nutrition Laboratory, Ithaca, New York 14583 Phone: (607)-255-5266. Fax: (607)-255-6683.

E-mail: jeanluc.jannink@ars.usda.gov

utility is also higher if crosses are only envisioned among the best parents rather than among all parents. Nevertheless, we show that among crosses the variance of progeny means is generally much greater than the variance of progeny standard deviations, restricting the utility of estimates of progeny standard deviations to a relatively small parameter space.

INTRODUCTION

In inbred line development, parents are crossed to generate segregating populations from which superior inbred progeny are selected. The value of a particular cross depends on the performance of its best progeny rather than on its mean progeny performance. In a typical breeding program, far too many crosses are possible between elite candidate parents for exhaustive evaluation. For example, among 50 elite parents there are 1225 possible crosses. Even if it were feasible to evaluate a sufficient set of progeny from all those crosses, it is unlikely that that would be efficient. Rather, one would want to predict, among possible crosses, which ones are most likely to lead to superior inbred lines.

SCHNELL and UTZ (1975) introduced the usefulness concept for line development. Their definition of the usefulness of the cross m was: $U_m = \mu_m + \Delta G_m = \mu_m + i\sigma_{G(m)}h_m$, where μ_m is the population mean of homozygous lines that can be derive from cross m, $\sigma^2_{G(m)}$ is the genetic variance among these lines, h_m is the square root of the heritability, and i is the standardized selection intensity. Two other criteria to similar usefulness are the varietal ability (WRIGHT, 1974; GALLAIS, 1979), and the probability of obtaining transgressive segregants (JINKS and POONI, 1976). Here, rather than focus on the genetic gain that might be obtained within a cross, we sought a simpler characterization that would express which crosses would generate progeny with higher genotypic values. Given the focus on genotypic value, we ignored the heritability to obtain what we call the superior progeny value, $s_m = \mu_m + i\sigma_{G(m)}$. With this definition, s_m equates to U_m with a heritability of 1.

In traditional breeding based solely on phenotypic measurements, μ_m can be predicted from the breeding values of the two parents but the only information available relevant to predicting $\sigma^2_{G(m)}$is the coancestry between parents. Thus, assuming two possible crosses have

identical μ_m, it is preferable to cross the parents with lower coancestries. After the advent of DNA markers, VAN BERLOO and STAM (1998) were the first to point out that marker information and quantitative trait loci (QTL) analysis could be used to identify complementary parents such that their progeny might segregate at more loci and show more extreme phenotypes. As in VAN BERLOO and STAM (1998), the breeding scenario investigated in this paper involves first deriving recombinant inbred lines (RIL) from a cross between two parents, then selecting among possible RIL pairs ones to cross to generate maximal superior progeny value. Without attempting to estimate a cross's σ_G^2, VAN BERLOO and STAM (1998), utilized a marker score computed from the flanking marker genotypes and weighted by QTL effects to discriminate among the crosses (VAN BERLOO and STAM, 1998).

More recently, BERNARDO et al. (2006) used QTL information to compute σ_G^2 to aid in the selection of crosses. In their computation, however, they assumed that the covariance between QTL effects could be ignored (BERNARDO et al. 2006), which is equivalent to assuming that all QTL resided on different chromosomes. As the ability to detect QTL improves and the number of QTL known to segregate within a population increases, however, accounting for linked QTL will become more important. In a toy example, we contrast Cross 1: [+ – +] × [– + –] with Cross 2: [+ + –] × [– – +], where + and – represent increasing and decreasing alleles. The variance among progeny from Cross 2 will be greater than that from Cross 1 because Cross 2 is more likely to generate progeny with [+ + +] and [– – –] genotypes that will have extreme phenotypic values. Thus, we need to account for recombination between QTL since two recombinations are required to generate those genotypes in Cross 1, but only one recombination in Cross 2.

The preceding discussion assumes previously-estimated QTL positions and effects. The method used to obtain these estimates, however, has a large impact on the effectiveness of marker-assisted selection (MAS) (HOSPITAL et al., 1997; MOREAU et al., 1998). The primary problem of QTL analysis is that the number of independent variables is large relative to the number of observations. Two different approaches have been used to deal with this situation, variable selection and shrinkage estimation.

Stepwise regression (JANSEN 1993; JANSEN and STAM 1994; KAO et al. 1999) is one

common procedure for variable selection in QTL analysis. A weakness of step-wise regression is that effects are included and removed from the model according to somewhat arbitrary statistical thresholds. Because many markers are tested in QTL mapping the process necessarily entails relatively high significance thresholds for marker inclusion in the model. A corollary is that included markers have inflated effect estimates (BEAVIS, 1994; SCHON et al., 2004; XU 2003a). On the other hand, the relaxed significance levels generally used for choosing significant markers for MAS (HOSPITAL et al., 1997; JOHNSON, 2001; BERNARDO et al., 2006), may lead to the inclusion of spurious markers. In the context relevant here of predicting a cross's mean and variance, both sorts of errors would be compounded.

New developments in shrinkage estimation seek to avoid variable selection by including all markers as predictors in the model and shrinking the allowed effect estimates toward zero, rather than choosing a "best" set among them. Ridge regression (HOERL and KENNARD, 1970) is a classical example of shrinkage estimation in which the least squares effect estimators $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ are replaced by $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$ (WHITTAKER et al., 2000). A high value for the parameter λ causes a penalty for large β thereby avoiding inflated estimates. This approach has strong affinities with the estimation of β using random Bayesian models that assumed a prior distribution for β: $\beta \sim N(0, \sigma^2_\beta)$.

A drawback of the ridge regression solution for including all markers is that all marker effects are equally penalized. To remove this constraint, XU (2003b) proposed a hierarchical model that allowed for a different variance for each β_i ($\sigma^2_{\beta_i}$), based on the random-model approach of MEUWISSEN et al. (2001). XU (2003b) showed that the posterior distributions of all parameters could be readily estimated using Markov chain Monte Carlo. His method performed well for both real and simulated datasets, though important improvements to the model were proposed by TER BRAAK et al. (2005). Because of the success of XU's model in QTL detection and the value of similar models in MAS (MEUWISSEN et al., 2001) we have adopted this approach in our analyses.

As presented thus far and as implemented in previous studies (e.g., BERNARDO et al., 2006), the prediction of superior progeny value is a multi-step analysis process. QTL analysis is first performed using one of the methods described above and the resulting map positions and effect estimates are then used to compute cross means and variances. We find fault with this two-step process because it prevents the individual or cross selection process from accounting for errors inherent to the QTL analysis. If, on the contrary, the selection process could account for the full uncertainty of the QTL analysis, different individuals or crosses might be selected. Bayesian analysis should allow MAS to account for uncertainty by using the full posterior distributions of the estimates of QTL effects.

The objectives of this study were first to specify more completely the theory to predict the value of a cross on the basis of its superior progenies, second to determine analytically the potential utility of accounting for the variance among a cross's progeny in predicting superior progeny value, and third to evaluate through simulation the effectiveness of different statistical approaches to predict superior progeny value. In particular, we wanted to contrast approaches that included or not an estimate of progeny variance in the prediction of superior progeny value; approaches that performed marker selection as opposed to including all markers in the QTL analysis; and approaches that split the QTL analysis from superior progeny value estimation in two steps as opposed to integrating them in a single step.

THEORY

Predicting the superior progeny value of a cross: As indicated above, for cross m, the superior progeny value s_m is, $s_m = \mu_m + i\sigma_{G(m)}$, and predicting it requires predicting μ_m and $\sigma_{G(m)}$ and defining a selection intensity, i. In what follows, we assume an additive model. Suppose there are L QTL affecting the phenotype in the whole population and L_m ($L_m \leq L$) loci segregating in cross m. Then the expected progeny value is a function of the L QTL effects and their genetic variance is a function of the segregating L_m QTL effects:

$$\mu_m = \underset{k}{E}\left(\sum_{i=1}^{L} Q_{ik(m)}\right) \qquad 1$$

$$\sigma^2_{G(m)} = \operatorname{var}_k\left(\sum_{i=1}^{L_m} sQ_{ik(m)}\right) \qquad 2$$

Where $Q_{ik(m)}$ is a random variable representing the effect of QTL i in progeny k of cross m, $sQ_{ik(m)}$ is a random variable representing the effect of segregating QTL i in progeny k of cross m. Note that if the parents of a cross carry the same allele at QTL, then the QTL will not segregate and $Q_{ik(m)}$ will be a constant. Expanding Equation 2 gives

$$\sigma^2_{G(m)} = \sum_{i=1}^{L_m} \operatorname{var}\left(sQ_{ik(m)}\right) + 2\sum_{i<j} \operatorname{cov}\left(sQ_{ik(m)}, sQ_{jk(m)}\right) \qquad 3$$

To calculate the terms in Equation 3, suppose the segregating QTL i and j recombine with rate c_{ij}, the homozygous effects of QTL i are $+\alpha_i$ and $-\alpha_i$ and those of QTL j are $+\alpha_j$ and $-\alpha_j$. Table 1 lists the inbred progeny frenquencies and genotypic values from a cross between a parent homozygous for the increasing allele at both loci and a parent homozygous for the decreasing allele at both loci (BULMER, 1985).

Given these frequencies and genotypic values,

$$\begin{aligned} \operatorname{var}\left(sQ_i\right) &= E\left(sQ_i^2\right) - \left[E\left(sQ_i\right)\right]^2 \\ &= \tfrac{1}{2}\left(+\alpha_i\right)^2 + \tfrac{1}{2}\left(-\alpha_i\right)^2 - 0 \\ &= \alpha_i^2 \end{aligned} \qquad 4$$

and

$$\begin{aligned} \operatorname{cov}\left(sQ_i, sQ_j\right) &= E\left(sQ_i sQ_j\right) - E\left(sQ_i\right)E\left(sQ_j\right) \\ &= \frac{0.5\alpha_i\alpha_j - c_{ij}\alpha_i\alpha_j - c_{ij}\alpha_i\alpha_j + 0.5\alpha_i\alpha_j}{1+2c_{ij}} \\ &= \frac{1-2c_{ij}}{1+2c_{ij}}\alpha_i\alpha_j \end{aligned} \qquad 5$$

Note that the covariance between QTL effects is positive in this case because the QTL were assumed in coupling in the parents crossed: one parent carried two increasing alleles while the other parent carried two decreasing alleles. To generalize across coupling and repulsion possibilities, the parameters $+\alpha_i$ and $+\alpha_j$ should be set to the QTL effects of one of the parents while $-\alpha_i$ and $-\alpha_j$ should be set to the QTL effects of the other parent. In this

way, the $\alpha_i\alpha_j$ term will be positive when QTL are in coupling and negative when they are in repulsion.

Substituting Equations 4 and 5 into Equation 3 gives

$$\sigma^2_{G(m)} = \sum_{i=1}^{L_m} \alpha_i^2 + 2\sum_{i<j} \frac{1-2c_{ij}}{1+2c_{ij}} \alpha_i \alpha_j .$$

Thus, predicting the genetic variance among inbred progeny of a cross between inbred parents requires estimates of homozygous QTL effects and of recombination frequencies between all pairs of QTL. Estimates of these parameters derive from the QTL analysis.

Utility of accounting for σ^2_G in predicting superior progeny value: The setup now is that two inbred lines that differ at *L* loci are crossed to generate a population of RIL. The objective then is to select pairs of RIL to cross to obtain maximal superior progeny value, *s*. We consider the variance of *s* and its origins. Given the definition $s_m = \mu_m + i\sigma_{G(m)}$ and assuming that μ and σ_G have zero covariance, $\text{var}(s) = \text{var}(\mu) + i^2\,\text{var}(\sigma_G)$. Thus, the influence of σ^2_G on *s* depends on the variance of μ relative to that of σ_G, and we investigate the ratio $t = \text{var}(\sigma_G) / \text{var}(\mu)$. Assume that QTL allele frequencies are 0.5, as would happen in a population derived from a cross between two inbred lines. For a single locus, three types of cross are possible between RIL from this population (Table 2).

If only a single QTL affects the trait in the population, then $\text{var}(\mu) = ½\alpha^2$ and $\text{var}(\sigma_G) = ¼\alpha^2$, such that $t = ½$. If *L* independent QTL affect the trait in the population, then $\mu = \sum_{i=1}^{L} Q_i$, where Q_i is the mean effect conferred by locus *i*, and

$$\text{var}(\mu) = \sum_{i=1}^{L} \text{var}(Q_i) = \frac{1}{2}\sum_{i=1}^{L} a_i^2 \qquad 6$$

For *L* independent loci, it is also simple to obtain $\text{var}(\sigma^2_G) = \sum_{i=1}^{L} \text{var}(\sigma^2_{Gi}) = \frac{1}{4}\sum_{i=1}^{L} \alpha_i^4$. Unfortunately, what we need is $\text{var}(\sigma_G)$. A first approach to obtain this variance is by the

delta method (LYNCH and WALSH, 1998). Using first order expansion, if $g(x)=\sqrt{x}$, then $\mathrm{var}\left[g(x)\right]=\mathrm{var}(x)g'\left[E(x)\right]=\frac{\mathrm{var}(x)}{4E(x)}$. Setting $x=\sigma_G^2$, we have

$$\mathrm{var}(\sigma_G)=\frac{\frac{1}{4}\sum_{i=1}^{L}\alpha_i^4}{4\left(\frac{1}{2}\sum_{i=1}^{L}\alpha_i^2\right)}=\frac{\sum_{i=1}^{L}\alpha_i^4}{8\sum_{i=1}^{L}\alpha_i^2} \qquad 7$$

Combining Equations 6 and 7 gives

$$t=\frac{\sum_{i=1}^{L}\alpha_i^4}{4\left(\sum_{i=1}^{L}\alpha_i^2\right)^2} \qquad 8$$

If all of the L loci have equal effects α, then the expression simplifies to $t = (4L)^{-1}$. Consequently as the number of independent loci of equal effect increases, the ratio t tends to zero and the influence of the variance of σ_G among crosses on superior progeny value becomes negligible. If the L loci do not have equal effects, but, as is often assumed (LANDE and THOMPSON, 1990), their variances follow a geometric series such that $\alpha_i^2=\alpha_{i-1}^2 a$, Equation 8 reduces to

$$t=\frac{1-a}{4(1+a)}=(4n_E)^{-1} \qquad 9$$

Where n_E is the effective number of QTL (LANDE and THOMPSON, 1990). Note that for $L = 1$, Equations 8 and 9 give $t = ¼$. We know, however, from the simple analysis of Table 2 that for a single-locus trait, $t = ½$. The discrepancy arises from the linear approximation used in the delta method to obtain Equations 8 and 9.

An exact expression for t assuming loci of equal effect that are unlinked and biallelic with allele frequencies of 0.5 can be obtained as follows. From Table 2, we know that the probability that a given cross will segregate at a given locus is 0.5. Assuming as before L independent QTL segregating in the population, then the probability that a given cross will segregate at L_m loci follows the binomial distribution $\binom{L}{L_m}0.5^{Lm}0.5^{L-Lm}=\binom{L}{L_m}0.5^{L}$. Given

loci of equal effect, the genetic variance generated from L_m loci will be $L_m\alpha^2$. Therefore, $E(\sigma_G) = \sum_{L_m=0}^{L} \binom{L}{L_m} 0.5^L \sqrt{L_m \alpha^2}$ and $[E(\sigma_G)]^2 = 0.5^{2L}\alpha^2 \left[\sum_{L_m=0}^{L} \binom{L}{L_m} \sqrt{L_m}\right]^2$. We thus obtain

$$\begin{aligned} \mathrm{var}(\sigma_G) &= E(\sigma_G^2) - [E(\sigma_G)]^2 \\ &= \frac{L}{2}\alpha^2 - 0.5^{2L}\alpha^2 \left[\sum_{L_m=0}^{L} \binom{L}{L_m} \sqrt{L_m}\right]^2 \end{aligned} \qquad 10$$

Combining Equation 10 with Equation 6 gives

$$\begin{aligned} t &= \frac{\frac{L}{2}\alpha^2 - 0.5^{2L}\alpha^2 \left[\sum_{L_m=0}^{L} \sqrt{L_m} \binom{L}{L_m}\right]^2}{\frac{L}{2}\alpha^2} \\ &= 1 - \frac{0.5^{2L-1}}{L}\left[\sum_{L_m=0}^{L} \sqrt{L_m} \binom{L}{L_m}\right]^2 \end{aligned} \qquad 11$$

Substituting $L = 1$ in Equation 11 does indeed give $t = ½$. Regardless of the approximation used, if QTL are independent, computing the ratio t shows that the influence of the variance among progeny within crosses on superior progeny value rather quickly becomes small (Figure 1). For example, with six unlinked QTL of equal or unequal variance, t is close to 1/20. The simulations of Figure 1 involved the following. A RIL population of 200 single seed descent progeny derived from a cross between two inbred lines was generated. For a given effective QTL number n_E, the rate of geometric decay of the variance was calculated as $a = (n_E - 1) / (n_E + 1)$, and the actual number of QTL simulated was twice n_E for n_E greater than five and ten for n_E less than or equal to five. In each simulation, the variances of μ and σ_G were calculated from 800 crosses chosen by randomly ordering the RIL into a loop then crossing each RIL with the four neighbors to either side of it. The ratio t was obtained as $t = \frac{1}{500}\sum_{j=1}^{500} t_j$ from 500 replicate simulations.

Because the simplifying assumption of independent loci rarely holds, we also assessed the impact of linkage on the ratio t through simulations similar to those for Figure 1. Instead

of being independent, QTL were randomly populated on one of the four different genomes: 5 chromosomes of 100 cM each, 10 chromosomes of 100 cM each, 20 chromosomes of 100 cM each, and 20 chromosomes of 200 cM each. The QTL variances were either equal or followed a geometric series. For each QTL, increasing and decreasing alleles were also assigned to parents at random.

From these simulations, we see that the effect of having a smaller genome is akin to the effect of having fewer QTL: the smaller the genome, the higher the ratio *t*, and the more relevant the variance of σ_G will be in determining superior progeny value (Figure 2). Nevertheless, the influence of this variance diminishes rather quickly with increasing QTL number (Figure 2). For example, for the genome with 10 chromosomes of 100 cM each, *t* is below 1/20 for 10 QTL. In general, then, when QTL number is high, accounting for σ_G will be of limited value. This was the phenomenon that BERNARDO et al. (2006) observed under the high QTL numbers that they simulated.

SIMULATIONS

Genetic model: The basic genetic model (Model A) for the population was as follows:

- Genomes were of ten chromosomes of 100 cM each and covered by markers every 10 cM.
- The genome was then populated with QTL at randomly chosen positions such that the effective QTL number n_E was 10. For each QTL, increasing and decreasing alleles were also assigned to parents at random. Thus coupling and repulsion linkages were generated at random. The QTL variances followed a geometric series (LANDE and THOMPSON, 1990).
- Genotypic values were calculated for 200 RIL progeny, and a normal deviate was added to the genotypic value to obtain phenotypic value assuming a heritability of 0.4.

A number of models that differed from the above in one parameter were tested, as follows.

Model B: Markers spaced every 20 cM rather than every 10 cM

Model C: Heritability of 0.1 rather than 0.4

Model D: Heritability of 0.8 rather than 0.4

Model E: Five rather than 10 effective QTL affected the trait

Model F: Twenty rather than 10 effective QTL affected the trait

Model G: Twenty rather than ten chromosomes

Model H: Chromosomes of 200 rather than 100 cM

Statistical analysis: The phenotypic values and marker information of the simulated RIL population were submitted to genome-wide Bayesian shrinkage analysis using the model proposed by XU (2003b) and implemented in WinBUGS (SPIEGELHALTER et al., 2007). Two chains were run, and after 5,000 burn-in iterations, 1,000 MCMC samples were thinned from a total of 20,000 iterations. Each sample consisted of the predicted genetic effects associated with all markers covering the genome. These data were used to obtain estimators of the superior progeny. For each estimator involving the among-progeny variance, the estimator was calculated for selection intensities of 20%, 15%, 10%, 5%, 2%, and 1%. Values of the standardized selection differential i corresponding to these intensities were calculated assuming progeny values were normally distributed. Six estimators were calculated as follows.

1. Full Bayesian treatment (denoted *sFull*). For MCMC sample j the superior progeny value of a cross m was calculated as ${}_j s_m = {}_j \mu_m + i\, {}_j \sigma_{G(m)}$ using sampled genetic effects for all markers. The estimator *sFull* was calculated as the mean sampled superior progeny value, $sFull = \frac{1}{1000} \sum_{j=1}^{1000} {}_j s_m$.

2. All marker posterior average treatment (denoted *sAll*). Average marker effects were calculated across all MCMC samples. For example, for marker i, $\alpha_i = \frac{1}{1000} \sum_{j=1}^{1000} {}_j \alpha_i$. Parameters $\overline{\mu}_m$ and $\overline{\sigma}_{G(m)}$ for a cross m were then calculated from these mean marker effects and $sAll = \overline{\mu}_m + i\overline{\sigma}_{G(m)}$.

3. All marker cross mean treatment (denoted μAll). Here simply $\mu All = \overline{\mu}_m$ from the *sAll* treatment.

4. Selected marker posterior average treatment (denoted *sSel*). Average marker effects were calculated as in *sAll*. Those markers that explained 2% or more of the total marker

variance were retained and used to calculate the parameters $\tilde{\mu}_m$ and $\tilde{\sigma}_{G(m)}$ for a cross *m*. Then, $sSel = \tilde{\mu}_m + i\tilde{\sigma}_{G(m)}$. This treatment most closely resembles a typical two-step approach of running QTL analysis first then using results of that analysis for MAS.

5. Selected marker cross mean treatment (denoted μSel). Here, $\mu Sel = \tilde{\mu}_m$ from the *sSel* treatment.

6. Phenotypic selection (denoted $\mu Phen$). The simplest approach used was to take the average phenotype of two parents as the prediction of their superior progeny mean.

These estimators of *s* were calculated for 800 random crosses chosen as in the ratio study above. To assess the utility of an estimator, we correlated it to the true superior progeny value calculated from the known simulated QTL effects and positions. For a given cross, the "true s_m" was calculated by simulating 5000 inbred progeny that might derive from it. The genotypic values of the top 20%, 15%, 10%, 5%, 2%, and 1% of these progeny were averaged and used as the true s_m for the corresponding selection intensity.

RESULTS

Under Model A the accuracy of estimators was $sFull > sAll > \mu All > sSel > \mu Sel > \mu Phen$ across all selection intensities (Figure 3a). While the inclusion of all markers in the model was more important than the inclusion of the term accounting for among-progeny variance, this latter term increased in importance as the selection intensity among progeny increased. The ordering changed when markers were spaced every 20 cM rather than every 10 cM (Figure 3b). The inclusion of all markers in the model remained far better than selecting markers before estimating superior progeny value, but with sparse markers, using estimates of σ_G to predict s_m appeared to introduce more error than information. Note that all estimators, save $\mu Phen$ that was not affected, were negatively affected by the decrease in marker density, though particularly those models incorporating the σ_G term suffered. The coarser marker grid presumably led to poorer estimation of the position of the QTL effects, which, in turn, affected estimates of σ_G. This result suggests that a marker spacing of 10 cM is minimal for this type of analysis and investigation of higher marker densities is warranted.

Under low heritability (Model C) the relative merit of the estimators involving markers was quite similar as under sparse markers: including all markers in the model was again the most important step to take, while incorporating estimates of σ_G made prediction worse (Figure 3c). It is also noteworthy that under the low heritability, even though only one or two QTL were correctly identified (data not shown), the prediction from μAll outperformed $\mu Phen$. Under high heritability (Model D), in contrast, σ_G was well-estimated and above a selection intensity of about 10%, all estimators that incorporated it did better than estimators that did not (Figure 3d). Interestingly also, at this high heritability the phenotype was such a good guide to the underlying genotypic value that $\mu Phen$ did better than μAll. For higher heritability, an index that incorporates phenotypic and marker information should be used to predict the cross mean (LANDE and THOMPSON, 1990). Once the cross mean is optimally predicted in that way, including consideration of among progeny variance might further prove valuable.

Given our previous analysis of the utility of including σ_G in the prediction of s_m, the impact of having few QTL (Model E) or many QTL (Model F) was not surprising. Under Model E, estimators that included σ_G were favored (Figure 3e), whereas under Model F they were penalized (Figure 3f). With few QTL, incorporating σ_G into the prediction had a greater beneficial effect than incorporating all markers (Figure 3e), contrary to the results found for the previous four models. In contrast, with many QTL, incorporating σ_G had a negative effect on prediction accuracy (Figure 3f). It may be that when more QTL are present, higher marker densities would be beneficial to tease them apart. In any event these simulations also make clear that with greater QTL numbers, less benefit should be expected from considering σ_G.

Finally, given the conditions of Model A, overall genome size and the allocation of the genome to many smaller chromosomes (Model G) or few larger chromosomes (Model H) did not affect the ranking of estimators (Figure 3a, 3g, and 3h). Results under the large genomes of Model G and H resembled each other and the results under Model A closely.

In the preceding simulation, we assessed the ability of the different estimators to discriminate between crosses among all progeny. In practice, breeders would not attempt crosses among all progeny but would only consider crosses among the best progeny (say,

those with high values). To evaluate the effect of considering crosses among only high-value progeny, we computed the correlation between the true and estimated s_m in Model A, using all 780 pair-wise crosses among the 40 RIL (out of 200) with the highest genetic values. In this case, incorporating σ_G into the prediction of s_m had an important beneficial effect that increased with the selection intensity (Figure 4). For randomly selected crosses, t was 0.04 (Figure 2b) but it increased to 0.21 for crosses among the best parents. Interestingly, for crosses among best parents, *μPhen* did better than either *μAll* or *μSel* (Figure 4), contrary to its behavior for crosses among all parents (Figure 3a).

DISCUSSION

Beyond results pertaining to specific genetic models, a number of results held across all the tested configurations. First, *μAll* was always superior to *μSel,* which means that avoiding model selection by including all markers in the final statistical model was always beneficial. This is consistent with other MAS studies (LANGE and WHITTAKER 2001; MEUWISSEN et al. 2001), which indicate that a better estimate of breeding values is obtained by incorporating all markers in the molecular score. Second, *sFull* always performed better than *sAll* (though often only slightly). Therefore, including the uncertainty of parameter estimation from QTL analysis appears always to be beneficial.

The fact that *μAll* outperformed *μPhen* at low heritability where few QTL were correctly identified (Figure 3c) indicates that genome-wide analysis models may capture at least a portion of the effects of QTL that they do not specifically identify. This phenomenon may have implications for how MAS statistical methods should deal with polygenic effects. These effects are typically included in models to account for loci of small effect that are not detected as QTL (KENNEDY et al., 1992). If statistical models including all markers capture variance from loci with very small effect, the polygenic effect may no longer be necessary. Indeed, two examples of MAS simulation exist where excellent response was obtained without a polygenic effect (MEUWISSEN et al. 2001; BERNARDO and YU, 2007). Whether this is a general phenomenon, or whether further improvement might be obtained by inclusion of a polygenic effect remains to be explored.

Both dense marker spacing and high heritability increased the accuracy of σ_G estimation due to the increased accuracy of marker effect and position estimation. Overall, it appears therefore that error in the estimates of marker effects, whether due to low heritability, sparse markers, or possibly small population size, has a more negative effect on the accuracy of estimates of σ_G than of μ. This fact, along with the generally low ratio of var(σ_G) to var(μ) limit the parameter space wherein it may be valuable to account for σ_G in the estimation of superior progeny value. Field experiments from different crop species also indicated that the usefulness of a cross is mainly influenced by the midparent value (GUMBER et al. 1999; UTZ et al. 2001; MIEDANER et al. 2006).

In our development, we assumed that μ and σ_G would have a covariance of zero. Intuitively, however, it seems unlikely that these parameters will be independent: two RIL that have similar extreme phenotypes (either high or low) may be fixed for the same alleles across a high fraction of loci. Thus, we would predict that extreme high or low μ will be associated with lower values of σ_G. In the general case, this mechanism would not generate a covariance between μ and σ_G, but in the case where crosses are only attempted between high-phenotype RIL (e.g., Figure 4), the mechanism will probably generate a negative covariance between the two. Nevertheless, we believe that the ratio between var(μ) and var(σ_G) that we have investigated will still be the most relevant single parameter to judge the utility of accounting for σ_G in making predictions.

The effect of considering crosses among only high-value progeny was primarily to decrease var(μ), which in turn enhanced the importance of accounting for var(σ_G) in the estimation of superior progeny value. The increase in the ratio t by a factor of 5.25 (from 0.04 to 0.21) can be attributed almost entirely to a drop in var(μ): under truncation selection with an intensity of 20%, the variance of the selected tail will be smaller by a factor of 4.05 relative to the variance of the distribution as a whole (FALCONER and MACKAY, 1997). The fact that t increased by more than that may indicate that truncation selection also increased var(σ_G), possibly because of negative linkage disequilibria among loci introduced by selection. The reason why $\mu Phen$ better predicted s_m than either μAll or μSel under these conditions is unclear. It may be that estimates of genotypic value derived from markers

decrease in accuracy as the genotypic value becomes more extreme. The phenotype, however, does not reflect the genotypic value less accurately at the extremes. We are not aware of previous reports of this phenomenon and if it indeed occurs it would warrant further investigation.

Another assumption that our setup forced was that allele frequencies in the initial population were 0.5. We briefly consider relaxing this assumption in the simplest way: if the favorable QTL allele frequency is p, the cross frequency row of Table 2 would become p^2, $2pq$, and q^2. Some algebra shows that $\text{var}(\mu) = 2pq\alpha^2$ whereas $\text{var}(\sigma_G) = 2pq\alpha^2(1-2pq)$ such that, for one QTL, $t = 1-2pq$. Thus, the ratio t is minimal for the case that we considered and, as p deviates from 0.5, t increases and accounting for σ_G may become more important.

While VAN BERLOO and STAM (1998) first presented the idea of using markers and QTL analysis to identify complementarity between parents, the simulations they presented did not directly assess whether using complementarity increased gain from selection relative to more standard MAS procedures. BERNARDO et al. (2006) found that estimating and accounting for σ_G in marker assisted recurrent selection generally did not lead to more rapid selection response (Table 2 of BERNARDO et al. 2006). Thus, their result is not in agreement with ours (Figure 4). Several differences in simulation conditions will have reduced the utility of accounting for σ_G in BERNARDO et al. (2006). First, their genome size (1746 cM) was greater and marker density (every 17 cM) was lower than presented here. In three out of four simulations, the number of individuals used in the QTL analysis (N = 100) was lower than here, which would have reduced accuracy of QTL estimation. Our results suggest that this accuracy is more critical to estimating σ_G than to estimating cross means (see for example, the effect of reduced heritability on the utility of σ_G, Figure 3). In addition, we simulated inbred lines while they simulated F_2 or S_0 lines, both of which provide less power and accuracy for QTL detection. Though they indicated that they generally detected about 40 QTL on a genome of 10 chromosomes, they did not account for QTL linkage in the calculation of σ_G, which would in principle lead to error in its prediction. Most importantly, however, three out of four of their simulation conditions involved either 40 or 100 QTL.

With these high QTL numbers we show that the ratio *t* would be very small such that, even absent errors in the QTL analysis, accounting for σ_G would be predicted to have low utility. There are nevertheless inconsistencies between their results and ours. For example, we would have predicted greater advantage to their "Unequal Fitness" methods (those that account for σ_G) in their genetic models with just 10 QTL. No trend in that sense was apparent. We also would have predicted greater advantage to the Unequal Fitness methods under high than low heritability. Again, no trend was apparent. We have no hypotheses to propose for the absence of these trends.

One aspect of MAS that we have emphasized here is the value of retaining information about the uncertainty of estimates from QTL analyses in the selection process. Indeed, the comparison of an estimator that did (*sFull*) versus did not (*sAll*) use the information showed that using it always improved the accuracy of estimates. Bayesian analysis, with its output of posterior distributions, facilitates the incorporation of uncertainty in analyses. Other studies on the value of crossing complementary parents have assumed that QTL information was known without error (HOSPITAL et al. 2000; SERVIN et al. 2004). HOSPITAL et al. (2000) used a recurrent selection framework in which the sole selection criterion depended on genotypes at markers flanking QTL. Complementation of QTL was introduced by measures to include parents carrying rare favorable QTL in the selected set. The study showed that the QTL complementation method was more efficient and robust than simple truncation selection on the marker score (HOSPITAL et al. 2000). SERVIN et al. (2004) took this approach one step further by considering an exhaustive list of possible pedigrees that could be used to pyramid a specified number of QTL. Given known QTL positions, the number of progeny required to generate the needed recombinants with a given probability at each generation can be calculated. In this way the process identifies the pedigree that can pyramid the QTL in a specified number of generations while requiring the evaluation of a minimum number of progeny. An important innovation brought by SERVIN et al. (2004) is that they consider a selection strategy planned over several generations whereas other MAS strategies operate one generation at a time (e.g., LANDE and THOMPSON, 1990; HOSPITAL et al. 2000; this study). The issue of optimal MAS considering an extended planning horizon was also

addressed by DEKKERS and VAN ARENDONK (1998), where the central issue was the appropriate weighting of QTL versus phenotypic information.

While HOSPITAL et al. (2000) and SERVIN et al. (2004) take a perspective that ignores the phenotype and is therefore quite different from the one adopted here, they also show that knowledge of marker segregation provides a benefit by allowing parents to be matched on a rational basis. The development of this "rational basis" has historically sought to tackle the problems of 1. How best to conduct the QTL analysis in view of the purpose of MAS (e.g., BERNARDO and YU, 2007); 2. How best to account for both QTL and phenotypic (or polygenic) information (e.g., LANDE and THOMPSON, 1990); 3. How to optimize plans over a horizon of longer than one generation (e.g., SERVIN et al. 2004); and 4. How to allow for other than additive modes of gene action (e.g, JANNINK, 2007). To these we add the question of considering error in QTL estimation. Clearly there remains a large terrain to explore in the combination of these five dimensions as they interact with the genetic determination of the trait(s) of interest. In addition, MAS methods must harmonize with plant breeding practice. For example, plant breeders usually generate many families each of relatively small size. Combining information from multiple families has been shown to be a powerful approach for QTL mapping (REBAÏ and GOFFINET, 1993; MURANTY 1996; XIE et al. 1998; XU, 1998; REBAÏ and GOFFINET, 2000; and VERHOEVEN et al., 2006; BLANC et al. 2006). Extending genome-wide MAS and the identification of complementary parents to this context should be valuable.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their comments and suggestions, which helped to improve the manuscript. This research was supported by USDA-NRI grant number 2003-35300-13202.

LITERATURE CITED

BEAVIS, W. D., 1994 The power and deceit of QTL experiments: lessons from comparative QTL studies, p. 250-265, In D. B. Wilkinson, ed. Proceedings of the 49th Annual Corn and Sorghum Research Conference. American Seed Trade Association, Washington, D.C.

BERNARDO R., MOREAU L., and A. CHARCOSSET, 2006 Number and fitness of selected individuals in marker-assisted and phenotypic recurrent selection. Crop Science **46**: 1972-1980.

BERNARDO, R., and J. YU, 2007 Prospects for genome-wide selection for quantitative traits in maize. Crop Science **47**: 1082-1090.

BLANC, G., A. CHARCOSSET, B. MANGIN, A. GALLAIS, and L. MOREAU, 2006 Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. Theor. Appl. Genet. **113**:206-224.

BULMER, M. G., 1985 The Mathematical Theory of Quantitative Genetics. Clarendon Press, Oxford.

DEKKERS, J. C. M., and J. A. M. VAN ARENDONK, 1998 Optimizing selection for quantitative traits with information on an identified locus in outbred populations. Genet. Res. 71:257–275.

FALCONER, D. S., and T. F. C. MACKAY, 1997 Introduction to Quantitative Genetics. Longman, New York.

GALLAIS, A., 1979 The concept of varietal ability in plant breeding. Euphytica **28**: 811–823.

GUMBER, R. K., B. SCHILL, and W. LINK, E. V. KITTLITZ, and A. E. MELCHINGER, 1999 Mean, genetic variance, and usefulness of selfing progenies from intra- and inter-pool crosses in faba beans (Vicia faba L.) and their prediction from parental parameters. Theor. Appl. Genet. **98**:569–580.

HOERL, A. E., and R. W. KENNARD, 1970 Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**: 55–67.

HOSPITAL, F., L. MOREAU, and F. LACOUDRE, A. CHARCOSSET, and A. GALLAIS, 1997 More on the efficiency of marker-assisted selection. Theor. Appl. Genet. **95**:1181–1189.

HOSPITAL, F., I. GOLDRINGER and S. OPENSHAW, 2000 Efficient marker-based recurrent selection for multiple quantitative trait. Genet. Res. **75**: 357–368.

JANNINK, J.-L., 2007 Identifying QTL by Genetic Background Interactions in Association Studies. Genetics **176**: 553-561.

JANSEN, R., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135**: 205-211.

JANSEN, R., and P. STAM, 1994 High resolution of quantitative traits into multiple loci via interval mapping. Genetics **136**: 1447–1455

JINKS, J.L., and H.S. POONI, 1976 Predicting the properties of recombinant inbred lines derived by single seed descent. Heredity **36**: 253–266.

JOHNSON, L., 2001 Marker assisted sweet corn breeding: A model for specialty crops. p. 25–30. In Proc. 56th Annu. Corn Sorghum Ind. Res. Conf. 5–7 Dec. 2001, Chicago, IL. Am. Seed Trade Assoc., Washington, DC.

KAO, C. H., Z-B. ZENG, and R. D. TEASDALE, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152**: 1203–1216.

KENNEDY, B. W., M. QUINTON, and J. A. M. VAN ARENDONK, 1992 Estimation of effects of single genes on quantitative traits. Journal of Animal Science **70**:2000-2012.

LANDE, R., and R. THOMPSON, 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics **124**:743–756.

LANGE, C., and J. WHITTAKER, 2001 On prediction of genetic values in marker-assisted selection. Genetics **159**: 1375–1381.

LYNCH, M., and B. WALSH, 1997 Genetics and Analysis of Quantitative Traits. Sinauer, Sunderland, Massachusetts.

MEUWISSEN, T. H. E., B. J. HAYES, and M. E. GODDARD, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**:1819-1829.

MIEDANER, T., B. SCHNEIDER and G. OETTLER, 2006 Means and variances for Fusarium head blight resistance of F2-derived bulks from winter triticale and winter wheat crosses. Euphytica **152**: 405–411.

MOREAU, L., A. CHARCOSSET, and F. HOSPITAL, and A. GALLAIS, 1998 Marker-assisted selection efficiency in populations of finite size. Genetics **148**:1353-1365.

MURANTY, H., 1996 Power of tests for quantitative trait loci detection using full-sib families in different schemes. Heredity **76**: 156–165.

REBAÏ, A. and B. GOFFINET, 1993. Power of tests for QTL detection using replicated progenies derived from a diallel cross. Theor. Appl. Genet. **86**: 1014-1022.

REBAÏ, A. and B. GOFFINET, 2000 More about quantitative trait locus mapping with diallel designs. Genet. Res. **75**: 243-247.

SCHON, C. C., H. F. UTZ, S. GROH, B. TRUBERG, S. OPENSHAW, *et al*., 2004 Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics **167**:485-498.

SERVIN, B., O. C. MARTIN, M. MEZARD, and F. HOSPITAL, 2004 Toward a theory of marker-assisted gene pyramiding. Genetics **168**:513–523.

SCHNELL, F.W., and H.F. UTZ, 1975 F1-Leistung und Elternwahl Euphy-der Züchtung von Selbstbefruchtern. p. 243–248. Bericht über die Arbeitstagung der Vereinigung österreichischer Pflanzenzüchter. BAL Gumpenstein, Gumpenstein, Austria.

SPIEGELHALTER, D. J., A. THOMAS, and N. G. M. BEST, 2007 WinBUGS Version 1.4 User Manual. Cambridge: Medical Research Council Biostatistics Unit. (Available from www.mrc-bsu.cam.ac.uk/bugs.)

TER BRAAK, C. J. F., M. P. BOER, and M. BINK, 2005 Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. Genetics **170**:1435-1438.

UTZ, H.F., M. BOHN, and A. E. MELCHINGER, 2001 Predicting progeny means and variances of winter wheat crosses from phenotypic values of their parents. Crop Science **41**:1470-1478

VERHOEVEN, K. J. F., J.-L. JANNINK, and L. M. MCINTYRE, 2006 Using mating designs to uncover QTL and the genetic architecture of complex traits. Heredity **96**:139-149.

VAN BERLOO, R. and P. STAM, 1998 Marker-assisted selection in autogamous

RIL populations: a simulation study. Theor. Appl. Genet. **96**:147–154

WANG, H., Y. M. ZHANG, X. M. LI, G. L. MASINDE, S. MOHAN, *et al*., 2005 Bayesian shrinkage estimation of quantitative trait loci parameters. Genetics **170**:465-480.

WHITTAKER, J. C., R. THOMPSON, and M. C. DENHAM, 2000 Marker-assisted selection using ridge regression. Genet. Res. **75**:249-252.

WRIGHT, A. J., 1974 A genetic theory of general varietal ability for diploid crops. Theor. Appl. Genet. **45**: 163—169.

XIE, C. Q., D. D. G. GESSLER and S. Z. XU, 1998 Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. Genetics **149**: 1139-1146.

XU, S., 1998 Mapping quantitative trait loci using multiple families of line crosses. Genetics **148**: 517-524.

XU, S., 2003a Theoretical Basis of the Beavis Effect. Genetics **165**:2259-2268.

XU, S., 2003b Estimating Polygenic Effects Using Markers of the Entire Genome. Genetics **163**:789-801.

Tables

Table 1. Inbred progeny frequencies and genotypic values from crossing a parent homozygous for the increasing allele with a parent homozygous for the decreasing allele at two loci. The loci recombine with frequency c_{ij} and inbred progeny are obtained by repeated generations of selfing.

Genotype	Progeny Frequency	Genotypic Value
$++$	$0.5 / (1 + 2\ c_{ij})$	$\alpha_i + \alpha_j$
$-+$	$c_{ij} / (1 + 2\ c_{ij})$	$-\alpha_i + \alpha_j$
$+-$	$c_{ij} / (1 + 2\ c_{ij})$	$\alpha_i - \alpha_j$
$--$	$0.5 / (1 + 2\ c_{ij})$	$-\alpha_i - \alpha_j$

Table 2. Three possible cross types and their frequencies assuming equal QTL allele frequencies. The genotypic value of the homozygous increasing allele is $+\alpha$ and that of the decreasing allele is $-\alpha$.

Cross Type	$[+] \times [+]$	$[+] \times [-]$	$[-] \times [-]$
Cross Frequency	0.25	0.50	0.25
μ	$+\alpha$	0	$-\alpha$
σ_G^2	0	α^2	0

Figures

Figure 1. Ratio *t* for independent QTL. Equal Exact: the ratio *t* for QTL with equal variances derived analytically. Unequal Simulation: the ratio *t* for QTL with geometrical-distributed variances derived from simulation. Delta Method: the ratio *t* for QTL with either equal or geometrical-distributed variances derived from Delta method.

Figure 2. Ratio *t* for different genome sets. Figure 2a represents simulation results with equal QTL variances. Figure 2b represents simulation results with QTL variances following a geometric series. 5Chr100cM: 5 chromosomes of 100 cM each; 10Chr100cM: 10 chromosomes of 100 cM each; 20Chr100cM: 20 chromosomes of 100 cM each; 20Chr200cM: 20 chromosomes of 200 cM each; Unlinked: independent QTL.

Figure 3. Correlations from random crosses between simulation truth and different predictors. Figure 3a-Figure 3h represent the results under Model A– Model H, respectively. Six selection intensity values, 1.40, 1.55, 1.76, 2.06, 2.42, and 2.67, correspond to the selection fraction of 20%, 15%, 10%, 5%, 2%, and 1%, respectively.

Figure 4. Correlations, corresponding to Model A, from top forty parent crosses between the simulation truth and different predictors.

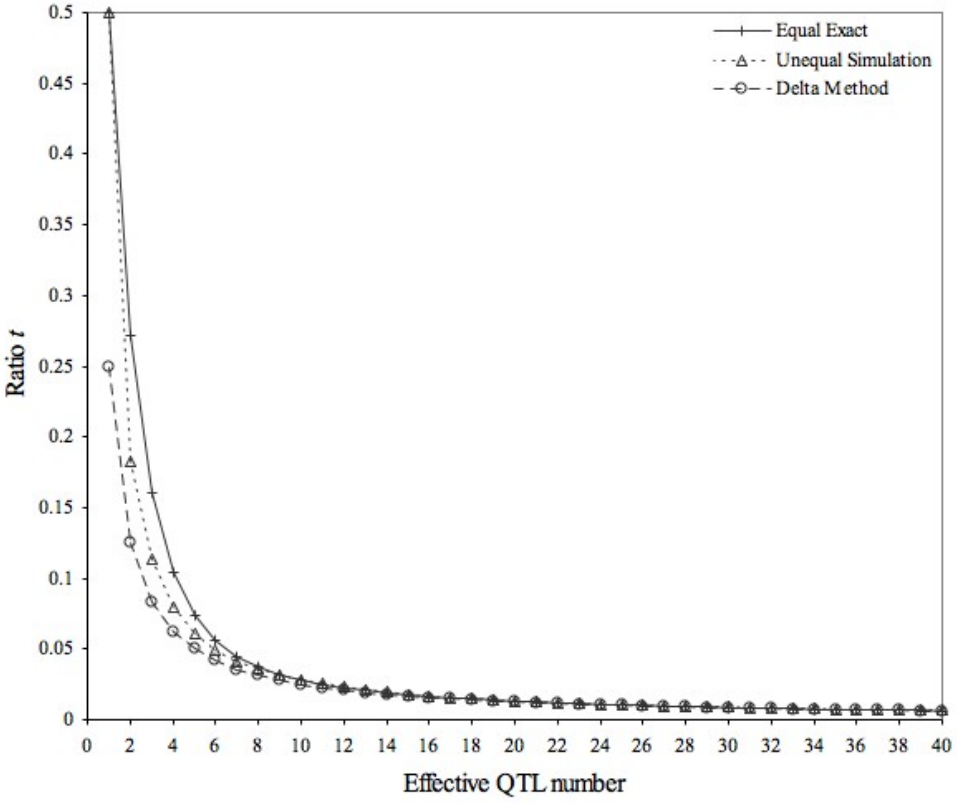
Equal Exact
Unequal Simulation
Delta Method
Ratio t
Effective QTL number
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40

Figure 1

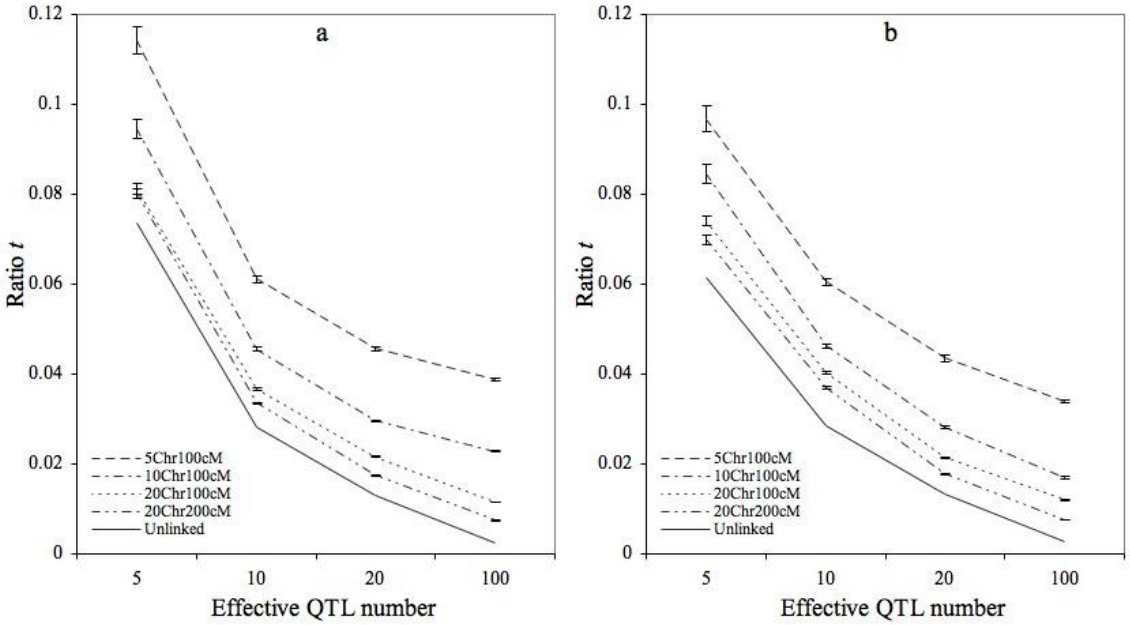
a
b
Ratio t
Effective QTL number
0.12
0.1
0.08
0.06
0.04
0.02
0
5
10
20
100
5Chr100cM
10Chr100cM
20Chr100cM
20Chr200cM
Unlinked

Figure 2

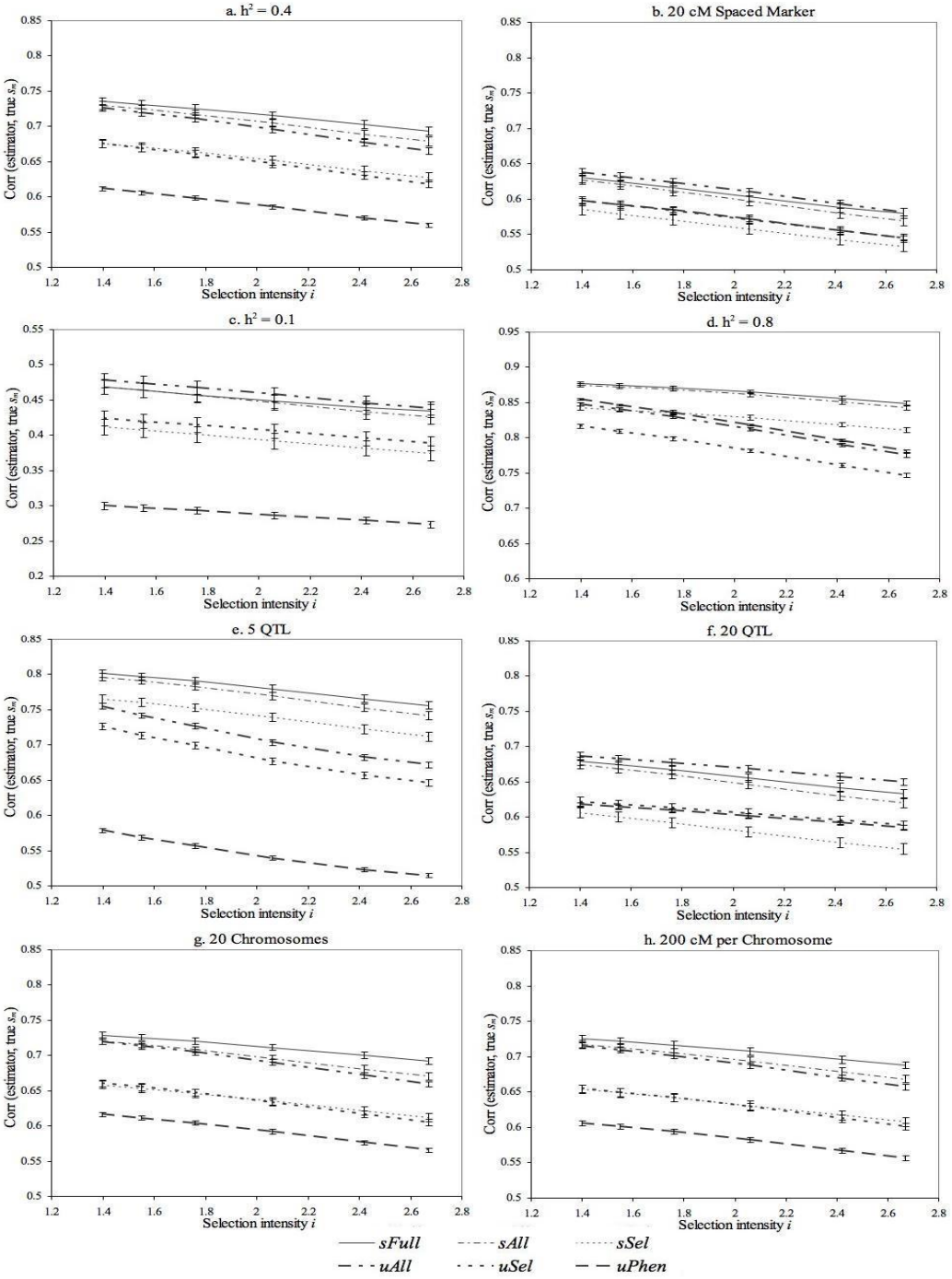
a. h² = 0.4
b. 20 cM Spaced Marker
c. h² = 0.1
d. h² = 0.8
e. 5 QTL
f. 20 QTL
g. 20 Chromosomes
h. 200 cM per Chromosome
Corr (estimator, true s_m)
Selection intensity *i*
sFull
sAll
sSel
uAll
uSel
uPhen

Figure 3

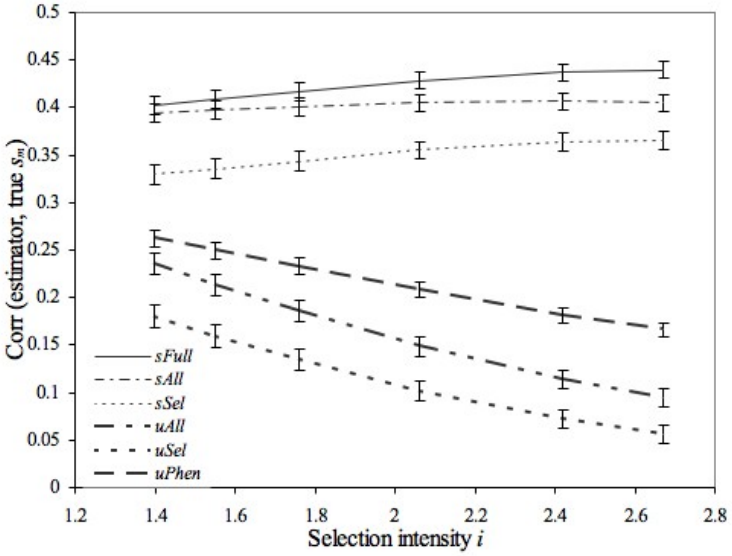
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Corr (estimator, true s_m)
sFull
sAll
sSel
uAll
uSel
uPhen
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
Selection intensity i

Figure 4

CHAPTER IV. ASSOCIATION-BASED GENOMIC SELECTION IN CULTIVATED BARLEY

A paper to be submitted to *Genetics*

Shengqiang Zhong, Jack C.M. Dekkers, and Jean-Luc Jannink

Abstract

Genomic selection, based on population-wide linkage disequilibrium (LD), has been pioneered in animal breeding. In genomic selection, all marker effects are first estimated on a training data set with marker genotypes and trait phenotypes. Breeding value can then be predicted for any genotyped individual in the population using all the estimated marker effects. There are relatively few studies in the plant breeding context. Simulations based on real barley SNP data were carried out to evaluate association-based genomic selection for cultivated barley. Forty-two spring two-row barley inbred lines were used as a starting germplasm pool. From these lines, two kinds of populations were generated to evaluate genomic selection: the first one was a typical plant breeding population with many families each of relatively small size; the second was a population derived after four generations of random mating of the original forty-two lines, which resulted in less LD. The first population had more within-family LD and higher population-wide LD. The performance of three genomic selection methods, random regression best linear unbiased prediction (RR-BLUP) assuming known equal marker variance, Bayesian shrinkage estimation, which shrinks small effects toward zero, and Bayes-B, for which the prior is that most markers have zero variance and few have large variance, were compared for these populations. A tradeoff exists between a method's ability to identify causal loci when the signal for those loci is strong (e.g., when causal locus allelic states are observed or strong LD exists) versus its ability to absorb the effect of loci when their signal is diffuse (e.g., the causal locus allelic states are not observed and LD is low). The analysis methods can be ordered according to this tradeoff such that the optimality of the method goes from the presence of specific signals to the presence of diffuse signals: Bayesian shrinkage estimation - Bayes-B - RR-BLUP. Overall, Bayes-B that fitted a

relatively higher proportion of markers into regression model compared to Bayes-B with fewer markers had more stable performance across scenarios with different QTL and marker numbers, size of the training dataset, and levels of linkage disequilibrium. Current barley SNP density around 500 SNP was found to be high enough for genomic selection to outperform phenotypic selection. Generally the prediction accuracy in the first population was better than that in the second population, which indicated that genomic selection can account within-family LD and population structure.

Introduction

Due to the advent of cheap and high density DNA chip technology, genome-wide dense molecular markers are becoming available for livestock and crop species. For example, the USDA sponsored Barley Coordinated Agricultural Project (CAP) is developing 3,000 informative SNP to be scored on 3,840 elite U.S. breeding lines. MEUWISSEN et al. (2001) developed a method called "genomic selection" to predict breeding values using genome-wide dense markers. Marker effects are first estimated on a training data set with marker genotypes and trait phenotypes. Breeding value can then be predicted for any genotyped individual in the population using the estimated marker effects. Studies have shown that genomic selection can lead to high correlations between predicted and true breeding value over several generations without repeated phenotyping (HABIER et al. 2007). Therefore, genomic selection can result in lower costs and increased rate of genetic gain. XU (2003) proposed a Bayesian shrinkage approach for QTL detection, based on the genomic selection idea (MEUWISSEN et al. 2001). TER BRAAK et al. (2005) proposed modifications to the XU (2003) model to ensure proper posterior distributions of marker variance and to better estimate QTL locations. ZHONG and JANNINK (2007) showed that breeding values predicted by the marker effects estimated by the Xu (2003) approach were useful for cross selection in inbred line development. We can regard the Bayesian shrinkage estimation by XU (2003) and TER BRAAK et al. (2005) as another form of genomic selection when using them to predict breeding values.

Genomic selection (MEUWISSEN et al. 2001) has been pioneered in animal breeding systems and few studies have been done in plant breeding systems. Genomic selection in plants has been studied for populations derived from bi-parental crosses in plant breeding (BERNARDO and YU, 2006; PIYASATIAN et al 2007; ZHONG and JANNINK, 2007). The success of genomic selection depends on the extent of and nature of linkage disequilibrium (LD) or the non-independence between marker alleles and quantitative trait locus (QTL) alleles in the population. When applied to bi-parental crosses, only within-family LD is used and thus effect estimates will be relevant only to that family. Genomic selection can also use population level LD that may exist across many parental lines, and then has the advantage of the broader inference of the whole breeding program. Plant breeding has some special characteristics relative to animal breeding. In particular, plant breeders often work with full-sib families created from crosses of inbred parents that vary in size, whereas half-sib families from non-inbred parents are more typical in animal breeding. Extensive LD will arise within each family but, given differing linkage phases across families, LD across a large set of families should represent the underlying population-wide LD. In typical plant breeding practice enough lines from a germplasm pool would be sampled such that associations with a dense marker set should be consistent across population, which would be particularly useful for association mapping or MAS in plant breeding programs. As an example of this idea, YU et al. (2008) proposed nested association mapping (NAM) to dissect quantitative traits in maize. In the NAM design, diverse inbred founders are selected and a large set of related progenies are generated for mapping. The founders have complete sequence or dense markers and the progeny genotypes are inferred through linkage using the sparse markers in the progenies. With this design, YU et al. (2008) successfully found a large fraction of the simulated QTL. In the present study, we are interested in genomic selection for designs that more closely approximate what may happen in breeding programs, in particular, designs that include larger numbers of smaller families than the designs investigated by YU et al. (2008).

Another characteristic of plant breeding is that the use of inbred lines is common and breeders usually have the ability to replicate individual genotypes over space and time and can thus obtain very accurate measurements of breeding values for a quantitative trait. Given a fixed amount of resources, breeders have the option to evaluate more individuals with

lower accuracy or fewer individuals with higher accuracy. These characteristics might affect how genomic selection should be carried out in crops relative to livestock.

Because of the particularities of crops, research into the use of genomic selection in crops is warranted. Barley offers an excellent public-sector model for self-pollinated crops due to the existence of a major effort in association genetics from the Barley coordinated agricultural project (CAP). A major goal of the Barley CAP is to provide accurate estimates of QTL effects from association mapping that can apply across all U.S. barley breeding programs. To date, the Barley CAP has generated close to 3000 SNP from expressed sequence tags (EST). These SNP have been scored on a core set of 102 inbred cultivated barley lines and genetic stocks, primarily of United States origin. Ninety-five of these lines have also been typed with DArT markers (WENZL et al. 2006). These lines include 42 two-row spring barley lines. In this paper, we use the structure and extent of LD of those 42 lines as a starting point to test genomic selection in a self-pollinating crop. In addition, we sought to evaluate the performance of different statistical models for genomic selection (MEUWISSEN *et al*. 2001; TER BRAAK et al. 2005), as affected by different breeding schemes. Evaluating these methods also gave us the opportunity to assess the adequacy of current barley SNP density for genomic selection, the effect of differing levels of within-family LD resulting from the breeding schemes, and the impact of different amounts of replication in the phenotypic evaluation of the breeding lines.

Materials and Methods

Germplasm and Map construction: To avoid excessive population structure due to the historical separation between two- and six-row barley, we chose to work only with the two-row set. Of the 102 Barley CAP Core lines that had both SNP and markers developed by diversity arrays technology (DArT, Wenzel et al 2006), 42 were two-row spring barley (Table 1). A total of 1803 mapped and polymorphic markers were scored on these lines.

The 1933-locus map constructed by Peter Szucs and Patrick Hayes on the Oregon Wolfe Barley population (http://www.barleycap.org/) was used as reference (it will be called the OWB map hereafter). This map contains DArT, SNP, and classic markers. The SNP were obtained from two Illumina Golden Gate oligonucleotide pool assays. One assay was

described in ROSTOKS et al. (2006) and the other was developed with similar methods. Map positions of SNP and classical markers based on the OWB map were obtained from HarvEST: Barley 1.64 (http://harvest.ucr.edu/). The SNP were obtained from the same pilot oligonucleotide assays as the OWB map; this map will be called the OPA map hereafter). A consensus map of DArT and classical markers was obtained from WENZL et al. (2006; called the DArT map hereafter). Because accurate map positions were not essential to the research presented here, the following expedient approach was used to merge these maps. For each chromosome, common markers between the OWB and each of the other two maps were identified. Per chromosome, there were on average 77 (range: 65 to 97) and 69 (range: 32 to 94) markers in common between the OWB and the OPA and DArT maps, respectively. The OWB map positions were regressed first on OPA, then on DArT map positions. Positions on the OPA and DArT maps were projected onto their OWB-predicted positions using these regressions. A marker's combined position was taken as the mean of all available OWB or OWB-predicted positions. From 1803 SNP, only those SNP that had minor allele frequency greater than 0.1 in the 42 two-row spring barley lines were used. This criterion resulted in the selection of 1605 SNP. In a second filter, SNP were chosen so that they were at least 0.2 cM or 0.75 cM apart, resulting in the selection of 1040 SNP and 575 SNP, respectively. To avoid dealing with missing data, missing SNP were randomly imputed according to their allele frequency. Since the rate of missing marker data was only 1.7%, the random imputation would not have large impact on final result.

Genetic model: The number of QTL was set at either 20 or 80. The trait was simulated with SNPs designated as QTL. Genetic variance caused by QTL followed a geometric series (LANDE AND THOMPSON 1990). QTL were assumed distributed at random across the genetic map. QTL locations were thus randomly drawn across the genetic map and the closest SNP to the chosen location was actually assigned to be a QTL. One SNP allele was randomly chosen to have a positive effect on the simulated trait, and the other SNP allele therefore had a negative effect on the trait. The size of the effect was scaled according to the SNP allele frequency to obtain the desired QTL variance for that SNP. The breeding value of a line was obtained by summing the effects of the QTL alleles that it carried and its phenotypic value by

adding to that a normal error deviate of a variance calculated to achieve the desired heritability.

Mating designs: Using the 42 lines as a founder pool, we evaluated genomic selection methods using four mating schemes.

Design 1: 42 families of 12 doubled haploid lines (DH) from the 42 two-rowed spring barley lines were generated using a single round-robin design (VERHOEVEN et al. 2005) in which 42 families were derived from the chain cross, i.e., inbred 1 x inbred 2, inbred 2 x inbred 3, ... , inbred 42 x inbred 1. From this design, 504 DH (= 42 x 12 DH per family) were used as the training dataset. Analyses were also performed on a "Double Design 1" where each family contained 24 DH, resulting in 1008 lines.

Design 2: 500 double haploids (DH) were derived after randomly mating the original 42 lines for four generations. The population size during random mating was set to 200 individuals. These 500 DH were then used as the training dataset. For comparison of different designs, the environmental variance was assumed to be constant for the first and second designs and was set such that the heritability was 0.4 for the original 42 lines. A Double Design 2 that contained 1000 DH was also analyzed.

Designs 3 and 4: The third and fourth mating schemes corresponded to replicating each evaluated line over more space and time than the previous two designs. The third design was similar to Design 1 and the forth design to Design 2. Instead of the family size of 12 in Design 1, a family size of 4 was used for the Design 3, which resulted in 168 lines in total. Instead of 500 lines in Design 2, only 168 lines were used for Design 4. The heritability for the last two designs was 0.67. This heritability was chosen given that we had one-third the number of lines and therefore could afford three times more replicates relative to the first two designs.

Regardless of the mating design used to generate the training dataset, 500 DH derived from random crosses among individuals in that training dataset were used as the testing dataset.

Linkage disequilibrium measures: Markers with minor allele frequency greater than 0.2 were used to estimate the extent of LD between all pairs of SNP within 100 cM in all chromosomes. The LD was computed as the squared correlation between alleles at two SNP,

$\hat{r}^2$ as the square of $\hat{r} = D_{ij} / \sqrt{p_i(1-p_i)p_j(1-p_j)}$ (HILL and ROBERTSON 1968), where $D_{ij} = p_{ij} - p_i p_j$, and p_{ij}, p_i, and p_j are the frequencies of haplotype *ij* and allele *i* at one locus and allele *j* at the other locus.

Statistical model: Three statistical models were used in this study to estimate genome-wide marker effects to estimate breeding values: two were as described by MEUWISSEN et al. (2001): random regression BLUP (RR–BLUP) and "Bayes-B", and the third was based on TER BRAAK et al.'s (2005) improvements to Bayesian shrinkage estimation (XU 2003), which shrinks small effects toward zero and we denote the "terBraak" approach. The RR-BLUP approach simply assumed that each marker has a variance equal to V_G/M, where V_G is the genetic variance and M is the marker number. In the Bayes-B approach, the prior of which for Bayesian analysis is that most markers have zero variance and few have large variance, MEUWISSEN et al. (2001) define π as the prior for the proportion of the markers associated with zero phenotypic variance, and assumed that it was known. In this paper, we evaluated two different values of π to allow the analysis to fit either a low or a high proportion of markers with non-zero phenotypic variance. The two values of π resulted in Bayes-B1 with π = (M–80)/M, and Bayes-B2 with π = (M–150)/M. Bayes-B2 therefore fitted a higher proportion of markers in the model than Bayes-B1. In some analyses, the causal QTL SNP were included along with all other markers. This inclusion represents an idealized case where QTL genotype is observed. Otherwise, the causal SNP was excluded, representing the typical case where the QTL is not observed. For each mating design and marker density scenario, 30 to 50 replicates were simulated and analyzed using these four methods. After each analysis, the estimated breeding values predicted for progeny in the testing dataset were correlated with their true breeding value known from the simulation. This prediction accuracy was used as the performance criterion for the methods.

Results

Extent of LD in two-row barley founders

Even when avoiding structure due to the division between two- and six-row barley, population-wide LD extended quite far (Figure 1, Figure 2A), in agreement with previous

studies in barley (KRAAKMAN et al., 2004; ROSTOKS et al., 2006). Long-range LD was much greater in our sample of two-row barley than expected in animal breeding, such as cattle breeding (ZENGER et al., 2007). Relative to the marker density of about 1 every cM available for this study, however, markers would rarely be in high LD with QTL: Even for QTL within 0.5 cM of a marker, the $\hat{r}^2$ was above 0.6 only about one eighth of the time (Figure 1). The single round-robin mating of Design 1 removed high LD ($\hat{r}^2$ greater than 0.4) at distances greater than 30 cM (Figure 2B), though moderate LD ($\hat{r}^2$ greater than 0.2) still extended to 100 cM. Moderate LD still occurred after the four rounds of random mating of Design 2 at distances of about 15 cM (Figure 2C). Design 2, however, eliminated long-distance LD above an $\hat{r}^2$ of 0.1. Recombination in Designs 1 and 2 greatly reduced long-distance LD, but, as expected, had little effect on LD at distances smaller than 2 cM (Figure 3). The average LD at 2 cM was around 0.2.

Prediction accuracy using genomic selection

Prediction accuracy, that is, the correlation between the breeding value predicted by genomic selection with the true value known from simulation, ranged from about 0.35 to 0.85 across the different scenarios analyzed (Figure 4). When the causal SNP were observed and QTL effects were large, the terBraak method gave the best, and RR-BLUP the worst, predictions (Figure 4A for 20 QTL). The performance of the terBraak method, however, declined sharply when either QTL effects were small (Figure 4A for 80 QTL) or when causal SNP were not observed (Figure 4B, 4C, and 4D). In almost all scenarios, Bayes-B2 outperformed Bayes-B1 (Figure 4B, 4C, and 4D); they performed equally only when there were 20 QTL and the causal SNP were observed (Figure 4A for 20 QTL). Likewise, in almost all scenarios, predictions were better when LD was high (Designs 1 and 3, Figure 4) than when it was low (Designs 2 and 4, Figure 4); the sole exception again being the scenario with 20 observed causal SNP. Predictions were more accurate in the dense than the sparse marker scenarios (Figure 4B versus Figure 4C), an effect that was accentuated under low as compared to high LD (Design 2 versus Design 1 in Figure 4B and 4C). The change in marker density did not, however, much affect the relative performance of the different analysis methods. Relative performance also changed little as a result of changes in the extent of LD

(Design 2 versus Design 1), although the terBraak method suffered the most from a decrease in LD, while RR-BLUP suffered the least, throughout. The RR-BLUP and Bayes-B2 methods performed quite similarly in all scenarios where causal SNP were not observed. Conditions that favored RR-BLUP over Bayes-B2 were when there were more QTL in the genetic model (i.e., 80 versus 20 QTL, for example, Figure 4B), when there were fewer markers in the analysis (i.e., 575 versus 1040 markers genome-wide, for example Figure 4C versus Figure 4B), when the training dataset was closer to being in linkage equilibrium (e.g., Design 2 versus Design 1, for example, Figure 4B), and when there were fewer individuals in the training dataset (i.e., 168 in Designs 3 and 4 versus 504 in Designs 1 and 2). Finally, predictions were generally more accurate for Designs 3 and 4, where fewer lines were phenotyped with more replication, than Designs 1 and 2 where more lines were phenotyped with less replication (Figure 4C versus Figure 4D). This effect interacted with the extent of LD; with high LD (Designs 1 and 3) prediction accuracies were similar, but with low LD (Designs 2 and 4) they were noticeably different (Figure 4C versus Figure 4D).

When causal SNP were observed, increasing the number of lines in the training dataset greatly increased prediction accuracy (Figure 5A). The terBraak method benefited the most from the increased amount of observations, while RR-BLUP benefited the least. Differences between methods in the effect of increasing size of the training data were great enough to cause some rank change in performance of the methods: whereas RR-BLUP outperformed Bayes-B1 when 500 DH were analyzed, the reverse was true when 1000 DH were analyzed (Fig 5A). When causal SNP were not observed, there was surprisingly little benefit to increasing the number of DH analyzed (Figure 5B). Almost no change in performance between Design 1 and Double Design 1 could be detected, while the improvement in performance of Double Design 2 over Design 2 was small, amounting to an increase in accuracy on the order of 0.02 to 0.04.

Discussion

Results showing high long-distance LD should not be surprising for this sample of two-row barley as it contains primarily North American but also European and Australian lines. It seems therefore likely that further population structure exists within our founder sample that

could generate LD between unlinked loci. The decline in LD due to intermating in Designs 1 and 2 was also as expected and generated a useful gradient of LD conditions upon which to evaluate the analysis methods.

Considering that most QTL would probably not have been in high LD with an observed SNP (Figure 1), genomic selection must be counted as a success: the baseline accuracy to which these methods should be compared is phenotypic selection. In that case, progeny performance is predicted by the average phenotype of the two parents. This correlation of mid-parent to single offspring is expected to be the square root of heritability times one-half, $\sqrt{\frac{1}{2}h^2}$ (FALCONER AND MACKAY 1996). For Designs 1 and 2, that comes to 0.45 for $h^2 = 0.4$: all methods of genomic selection out-performed phenotypic selection except terBraak approach on Design 3 (Figure 4B and 4C). For Designs 3 and 4, $\sqrt{\frac{1}{2}h^2} = 0.58$ which is lower than the accuracies of Bayes-B2 and RR-BLUP on Design 3. Assuming 80 QTL and unobserved causal SNP, the best methods of analysis provide accuracies that are equal to phenotypic selection with a heritability of 0.77 for Design 1 and 0.61 for Design 2.

A general interpretive scheme that we believe explains many of the patterns of relative performance among the analysis methods is that a tradeoff exists between a method's ability to identify causal loci when the signal for those loci is strong (e.g., when causal locus allelic states are observed or when strong LD exists) versus its ability to absorb the effect of loci when their signal is diffuse (e.g., the causal locus allelic states are not observed and LD is low). The analysis methods can be ordered according to this tradeoff such that the optimality of the method goes from the presence of specific signals to the presence of diffuse signals: terBraak - Bayes-B1 – Bayes-B2 - RR-BLUP. In our analyses, the strongest locus-specific signals occurred when there were few (20) large QTL and the SNP causing their effect was observed (Figure 4A). In that case, terBraak performed best and RR-BLUP worst. Conversely such signals were most diffuse when there were many (80) small QTL, markers were sparse, and LD was diminished by random mating (Figure 4D). In that case, terBraak performed worst and RR-BLUP best. HABIER et al. (2007) identified two components contributing to prediction accuracy in genomic selection, one due to LD between markers

and QTL and the other due to genetic relationships between individuals that can be captured in the absence of LD. The distinction we make here between specific and diffuse signals is perhaps identical though we stress also the importance of QTL effect size in identifying signal through LD.

The tradeoff between ability to capture LD versus genetic relationship signals is not absolute. For example, Bayes-B2 and Bayes-B1 were essentially equal in their ability to capture the LD signal of 20 observed QTL (Figure 4A) but Bayes-B2 was superior to Bayes-B1 in capturing the genetic relationship signal when LD was weak (Figure 4C). The superiority of Bayes-B2 over Bayes-B1 in estimating genetic relationships came from the fact that it fit more markers in the model (HABIER et al., 2007), as shown in Figure 6. The terBraak method was least able to capture genetic relationships even though, just as for RR-BLUP, it maintains all markers in the model. The terBraak model, however, severely shrinks the effects of markers that are only weakly related with the phenotypes, such that their weight in estimating genetic relationships is practically null. To compensate for this weakness in the terBraak method, it would be possible to add a polygenic effect into the model. One concern is that the polygenic effect will be confounded with SNP effects. In one study, including a polygenic effect in genomic selection models hardly affected prediction accuracy (CALUS and VEERKAMP, 2007).

The simplicity of RR-BLUP also seemed to enhance its prediction accuracy when few observations were available to estimate marker effects. Designs 3 and 4 had fewer individuals available for analysis than Designs 1 and 2 (168 for Designs 3 and 4 as compared to 504 and 500 for Designs 1 and 2, respectively), and resulted in RR-BLUP being best among all methods. This observation suggests that the Bayes-B and terBraak methods have greater data requirements than does RR-BLUP. For Bayes-B, if there is insufficient data to allow the model to determine whether a given marker should be in or out of the model, its more complex prior specification of the QTL variance than RR-BLUP will not improve model performance. Lack of data would be evidenced by only small differences among markers in their probability of being included in the model. Increasing the number of observations clearly increased these differences (Figure 6A versus 6B), indicating that the relative performances of Bayes-B and RR-BLUP will likely be sensitive to the amount of

data available. Similarly, the terBraak method would likely need to have a higher data requirement than RR-BLUP.

The ability to capture QTL effects through LD with markers was clearly important, given that in almost all cases prediction accuracies were greater when LD was high (under Design 1) than when it was low (under Design 2). The sole exception to this trend was for the terBraak and Bayes-B analyses when the causal SNP were observed (Figure 4). Here it may be that these MCMC analyses had more difficulty converging when markers were in high LD (TER BRAAK et al. 2005). The decline in prediction accuracy for RR-BLUP as a result of lower LD in this case (Figure 4A) may have been because, even when the causal SNP was observed, RR-BLUP relied on more than one marker to absorb the full QTL effect. The greater penalty placed by RR-BLUP on large QTL effects due to its assumption that QTL variances were equal to V_G / M may have prevented RR-BLUP from effectively capturing QTL effects even when the QTL was typed. It was not clear *a priori* whether the RR-BLUP assumption that all marker variances are equal would cause it to perform better under high or low LD. Indeed, LD extends far, one might assume that all markers are in LD with one or more QTL, and therefore that the assumption fits well. Careful examination of Figure 4 however, shows that RR-BLUP performed better relative to other methods when LD was low than high, even though it seems likely that with low LD the assumption of equal marker variances will be further from the truth. This is consistent with what HABIER et al. (2008) demonstrated: LE markers can capture the genetic relationship and RR-BLUP is better than Bayes-B in this regard. BERNARDO and YU (2007) investigated genomic selection for DH lines generated from a single bi-parental cross, which would have long-range LD. They found that the genetic gain when assuming equal variance for all markers, as in RR-BLUP, was similar to BLUP using the true variances for the markers. It seems likely that the performance of BLUP using the true variance for markers would be an upper limit on methods such as Bayes-B that seek to relax the assumption of equal variance.

Not surprisingly, there was a clear interaction between the extent of LD and marker density, such that at high density, lower LD caused a smaller drop in prediction accuracy (Figure 4B and 4C) and, by the same token, at high LD, a decrease in marker number hardly affected performance (Figure 4B and 4C). This observation underscores the importance of

knowing the extent and structure of LD in determining requisite marker densities.

Especially at low LD, evaluating fewer individuals more extensively improved prediction accuracy (Figure 4D). It is well known that improvements in capturing QTL effects through LD can be obtained by allocating observations to more distinct genotypes than to allocating observations to replications of a subset of genotypes (KNAPP and BRIDGES, 1990). The improvement in accuracy apparent in Designs 3 and 4 relative to Designs 1 and 2 must therefore have come simply from improvements in the accuracy of the contribution of the genetic relationship to the prediction due to the greater heritability of observations in Designs 3 and 4. Further investigation of genomic selection in later generations, using the approach of HABIER et al. (2008), will verify whether the improved accuracy is from the contribution of the genetic relationship.

Keeping everything else constant, increasing the number of observations from 500 to 1000 increased prediction accuracy (Figure 5). Surprisingly, this increase was most noticeable when the QTL were typed than when they were unobserved (Figure 5A versus 5B). We have sought to explain the performance patterns of the analysis methods according to their ability to capture QTL effects through LD and through average genetic relationships between lines. Of these two, increasing observation number should improve accuracy due to the LD component, but not due to the genetic relationship component. The fact that accuracy did not improve suggests that the accuracy is due mostly to the genetic relationship component. That hypothesis is supported by the evidence from the posterior probabilities of including markers into the model (Figure 6) that show increases in the probabilities of causal SNP, while there were no obvious increased posterior probabilities (Figure 7) for markers that had the highest LD with causal SNP when causal SNP were untyped. Increasing the population size to 2000 significantly increased the posterior probabilities of markers in highest LD when causal SNP were untyped (result not shown). Therefore this further supports the idea that the data requirement to capture QTL effects by LD for genomic selection is greater than to capture overall genetic relationships. If the sample size is small, the prediction accuracy from Bayes-B would likely mostly come from the genetic relationship component. Again the prediction accuracy in further generations should show this.

In conclusion, the SNP density is high enough for genomic selection to select individuals with the highest breeding values. Generally the prediction accuracy in Design 1 was better than in Design 2, which indicated that Bayes-B could account for within-family LD and population structure. Overall, Bayes-B2 had the most stable performance across different scenarios.

Literature Cited

BERNARDO, R., AND J. YU, 2007 Prospects for genome-wide selection for quantitative traits in maize. Crop Sci. 47: 1082–1090.

CALUS M. P. L., R.F. VEERKAMP 2007 Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. Journal of animal breeding and genetics 124: 362–368.

FALCONER, D. S., and T. F. C. MACKAY, 1997 Introduction to Quantitative Genetics. Longman, New York.

Habier, D., R. L. Fernando and J. C. M. Dekkers. 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics. 177(4):2389-97.

HALEY C.S., and P.M. VISSCHER 1998 Strategies to utilize marker-quantitative trait loci associations. J. Dairy Sci., 81(Suppl. 2), 85–97.

HILL, W. G. and A. ROBERTSON 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38, 226-231.

KNAPP, S.J., AND W.C. BRIDGES. 1990 Using Molecular Markers to Estimate Quantitative Trait Locus Parameters: Power and Genetic Variances for Unreplicated and Replicated Progeny. Genetics 126:769-777.

KRAAKMAN, A.T.W., R.E. NIKS, P.M.M.M. VAN DEN BERG, P. STAM, and F.A. VAN EEUWIJK. 2004 Linkage Disequilibrium Mapping of Yield and Yield Stability in Modern Spring Barley Cultivars. Genetics 168:435-446.

LANDE, R., and R. THOMPSON, 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756.

MEUWISSEN, T. H. E., B. J. HAYES, AND M. E. GODDARD, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

PIYASATIAN, N., R. L. FERNANDO, and J. C. M. DEKKERS, 2007 Genomic selection for marker-assisted improvement in line crosses. Theor. Appl. Genet. 115:665-674.

ROSTOKS, N., L. RAMSAY, K. MACKENZIE, L. CARDLE, P. R. BHAT et al., 2006 Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proceedings of the National Academy of Sciences 103: 18656-18661.

TER BRAAK, C. J. F., M. P. BOER, and M. BINK, 2005 Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. Genetics 170:1435-1438.

VERHOEVEN, K. J. F., J.-L. JANNINK, and L. M. MCINTYRE, 2006 Using mating designs to uncover QTL and the genetic architecture of complex traits. Heredity **96**:139-149.

WENZL, P., H. LI, J. CARLING, M. ZHOU, H. RAMAN et al., 2006 A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits. BMC GENOMICS 7: Article 206.

XU, S., 2003 Estimating Polygenic Effects Using Markers of the Entire Genome. Genetics 163:789-801.

YU J, HOLLAND J B, M D MCMULLEN, and E S. BUCKLER 2008 Genetic Design and Statistical Power of Nested Association Mapping in Maize. Genetics 178: 539-551.

ZENGER KR, KHATKAR MS, CAVANAGH JA, HAWKEN RJ, RAADSMA HW. 2007 Genome-wide genetic diversity of Holstein Friesian cattle reveals new insights into Australian global population variability, including impact of selection. Animal Genetics. 38: 7-14.

ZHONG, S., and J. JANNINK, 2007 Using QTL results to discriminate among crosses based on their progeny mean and variance. Genetics. 177:567-576.

Table 1. Two-row spring barley lines

Line	CAP core #	Variety (selection; ID)	Line	CAP core #	Variety (selection; ID)	Line	CAP core #	Variety (selection; ID)
1	12	B1202	15	32	Newdale	29	67	CIho 4196
2	13	CDC Kendall	16	33	TR306	30	68	Craft
3	14	Merit	17	34	C-14	31	69	Hockett
4	15	Klages	18	35	Franklin	32	70	Geraldine
5	16	B1215	19	36	Pasadena	33	71	Eslick
6	17	Garnett	20	37	Flagship	34	72	Haxby
7	18	CDC Stratus	21	41	Orca	35	73	Hays
8	19	AC Metcalfe	22	42	BCD47	36	80	Sublette
9	20	Baronesse	23	50	Conlon	37	83	Radiant
10	22	Arapiles	24	51	ND21863	38	84	Crest
11	23	Collins	25	52	Rawson (ND19119-2)	39	85	Farmington
12	24	Scarlett	26	53	Bowman	40	89	Conrad
13	30	Harrington	27	54	Shenmai 3	41	90	2B98-5312
14	31	CDC Copeland	28	55	Canela	42	91	2B96-5038

Figures

Figure 1. Frequency distribution of LD estimates, as measured by by $\hat{r}^2$, in a sample of 42 spring two-row barley cultivars, for all markers with minor allele frequency > 0.2.

Figure 2. Decline of LD as measured by $\hat{r}^2$ against distance in cM, for all markers with minor allele frequency > 0.2. A, B and C show $\hat{r}^2$ in the original forty-two lines, design 1 and design 2, respectively.

Figure 3. Decline of the moving-average LD of 0.25 cM interval for all markers with minor allele frequency > 0.2.

Figure 4. Correlation between simulated and predicted breeding values. Graphs A and B show analyses under dense marker cases. Graphs C and D show analyses under the sparse marker cases. Graph A shows analyses when the causal SNP has been typed while graphs B through D show results when the causal SNP is unobserved. The standard error for each point is very small (below 0.002) and is thus not shown. Note that the y-axis for graph A is on a different scale from the other graphs.

Figure 5. Prediction accuracy with different population sizes under sparse marker simulations. Results were averaged from 30 replicates; the standard errors were very small (below 0.002) and are thus not shown. All scenarios are under the 80 QTL condition. Graph A shows the analyses with observed causal SNP, and graph B with unobserved causal SNP. Note that the y-axis scale for graph A is different from that for graph B.

Figure 6. Posterior probabilities under observed causal SNP condition for the sparse marker case. A is for Design 2. B is for Double Design 2. The posterior probabilities for causal SNP for Bayes-B2 are circled.

Figure 7. Posterior probabilities under unobserved causal SNP condition for the sparse marker case. A is for Design 2. B is for Double Design 2. The posterior probabilities for the markers that have highest LD with causal SNP for Bayes-B2 are circled.

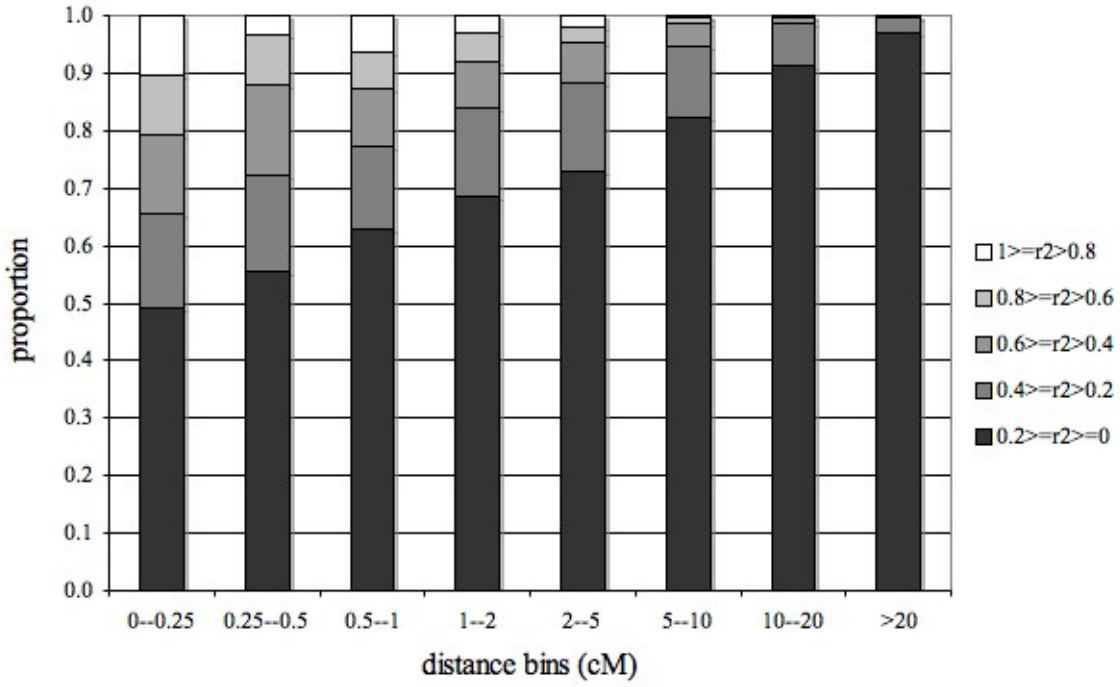
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
proportion
0--0.25
0.25--0.5
0.5--1
1--2
2--5
5--10
10--20
>20
distance bins (cM)
1>=r2>0.8
0.8>=r2>0.6
0.6>=r2>0.4
0.4>=r2>0.2
0.2>=r2>=0

Figure 1

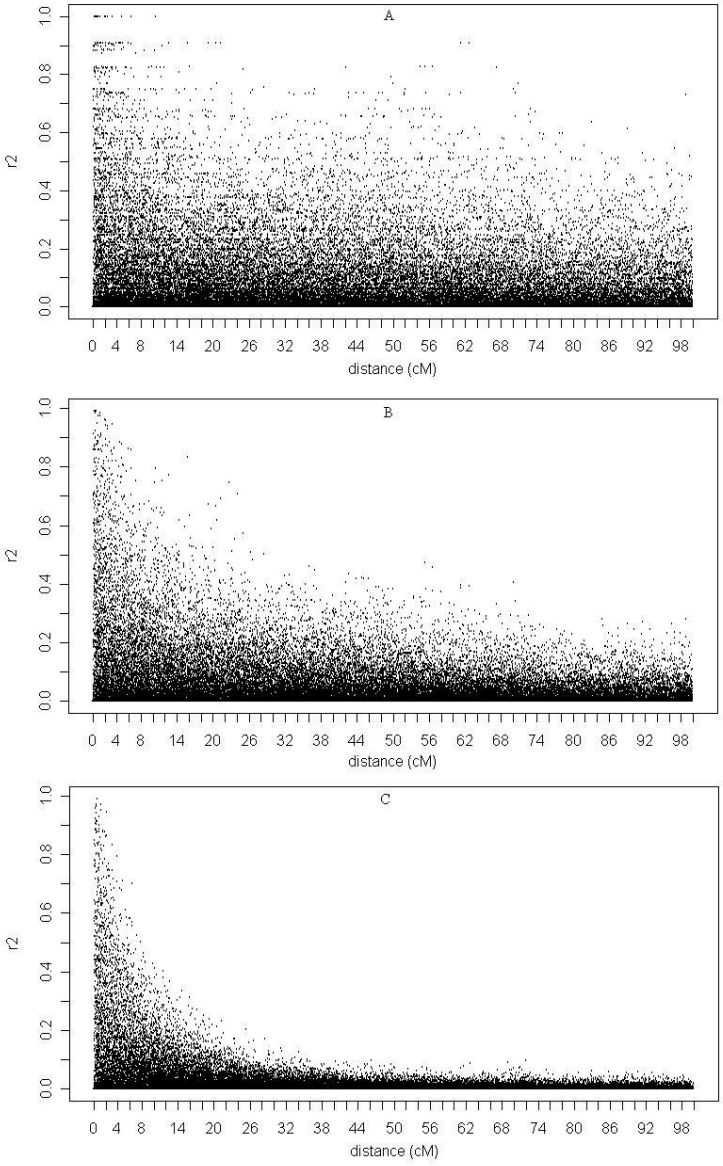
A
B
C
r2
0.0
0.2
0.4
0.6
0.8
1.0
0 4 8 14 20 26 32 38 44 50 56 62 68 74 80 86 92 98
distance (cM)

Figure 2

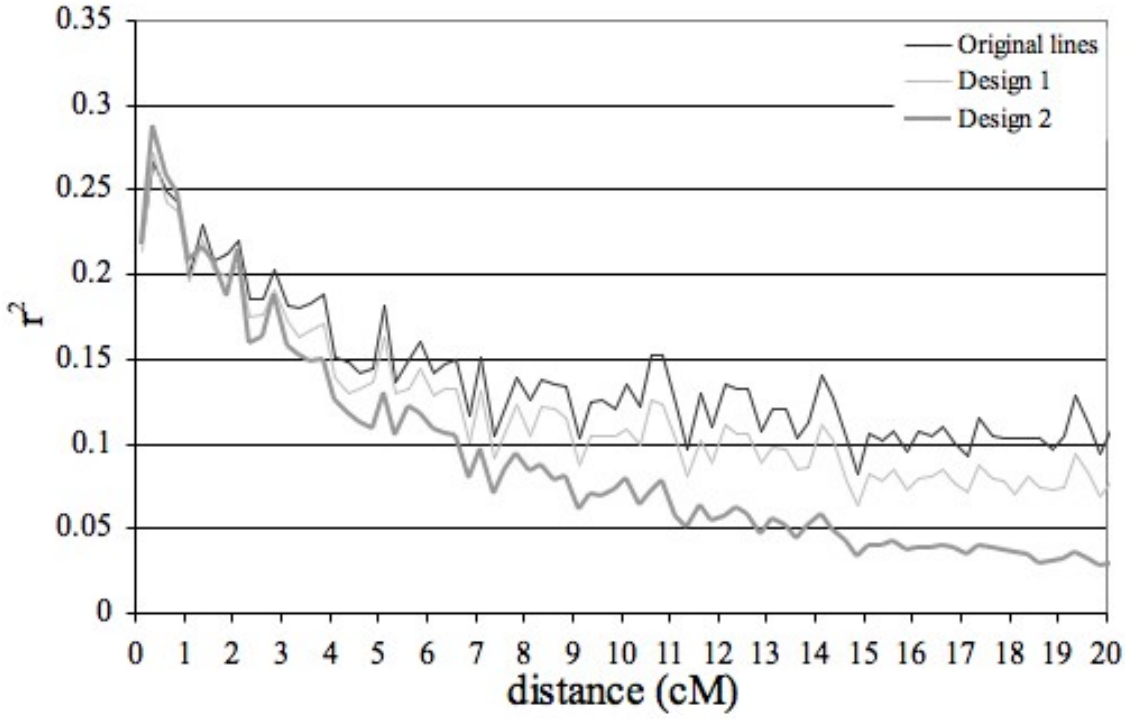
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
r2
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
distance (cM)
Original lines
Design 1
Design 2

Figure 3

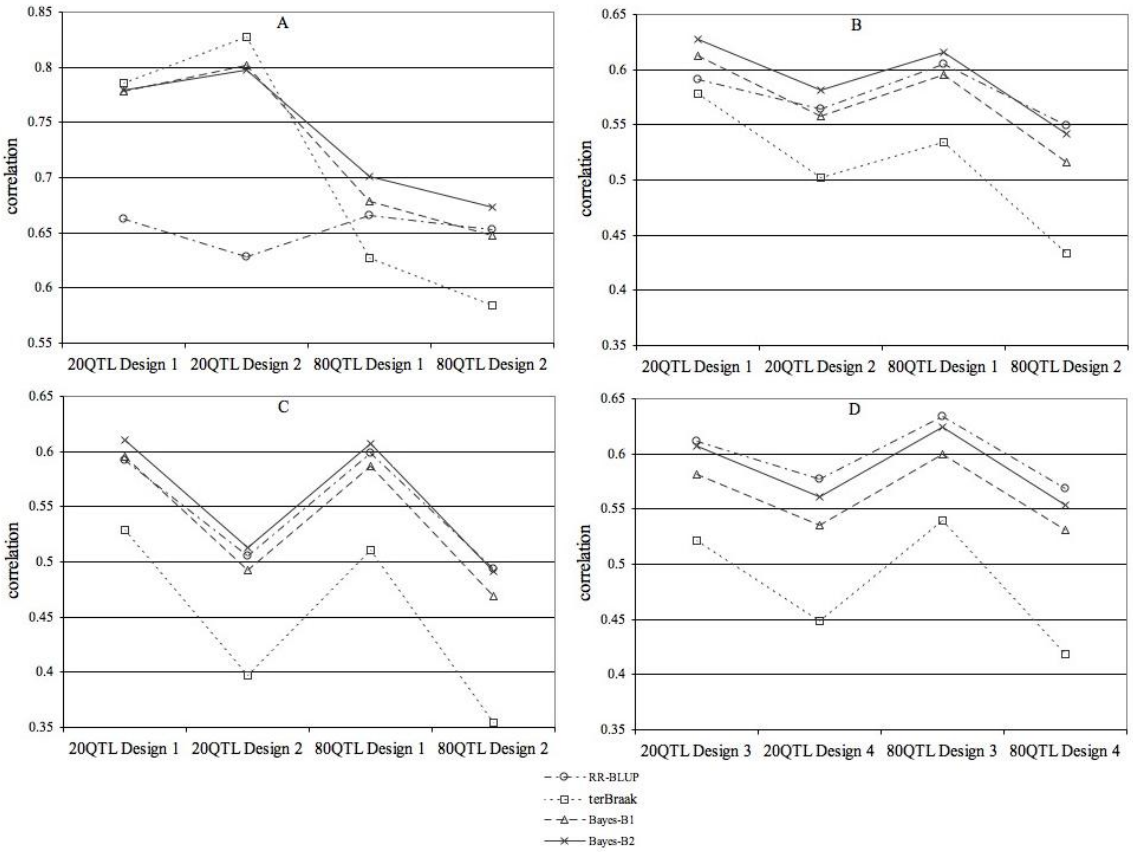
A
B
C
D
correlation
20QTL Design 1
20QTL Design 2
80QTL Design 1
80QTL Design 2
20QTL Design 3
20QTL Design 4
80QTL Design 3
80QTL Design 4
RR-BLUP
terBraak
Bayes-B1
Bayes-B2

Figure 4

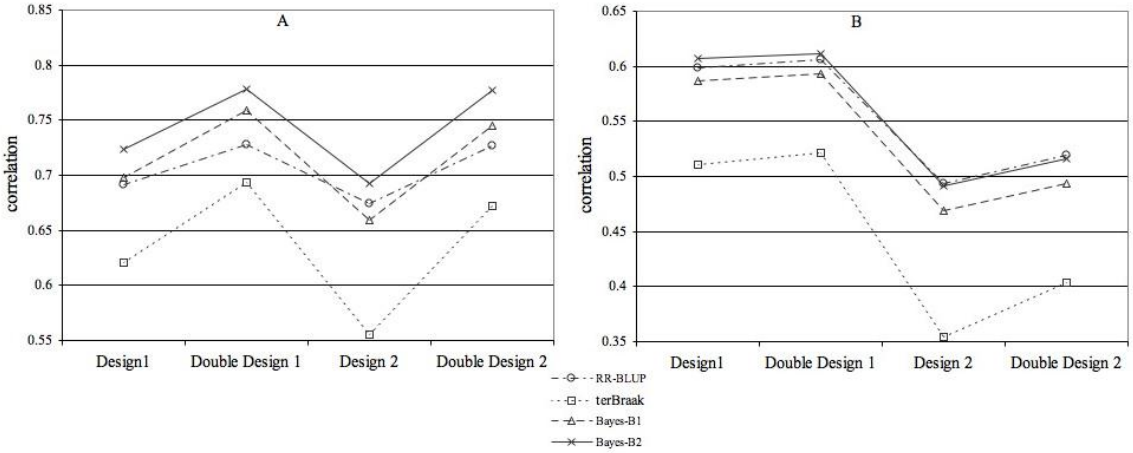
A
B
correlation
correlation
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.65
0.6
0.55
0.5
0.45
0.4
0.35
Design1
Double Design 1
Design 2
Double Design 2
RR-BLUP
terBraak
Bayes-B1
Bayes-B2

Figure 5

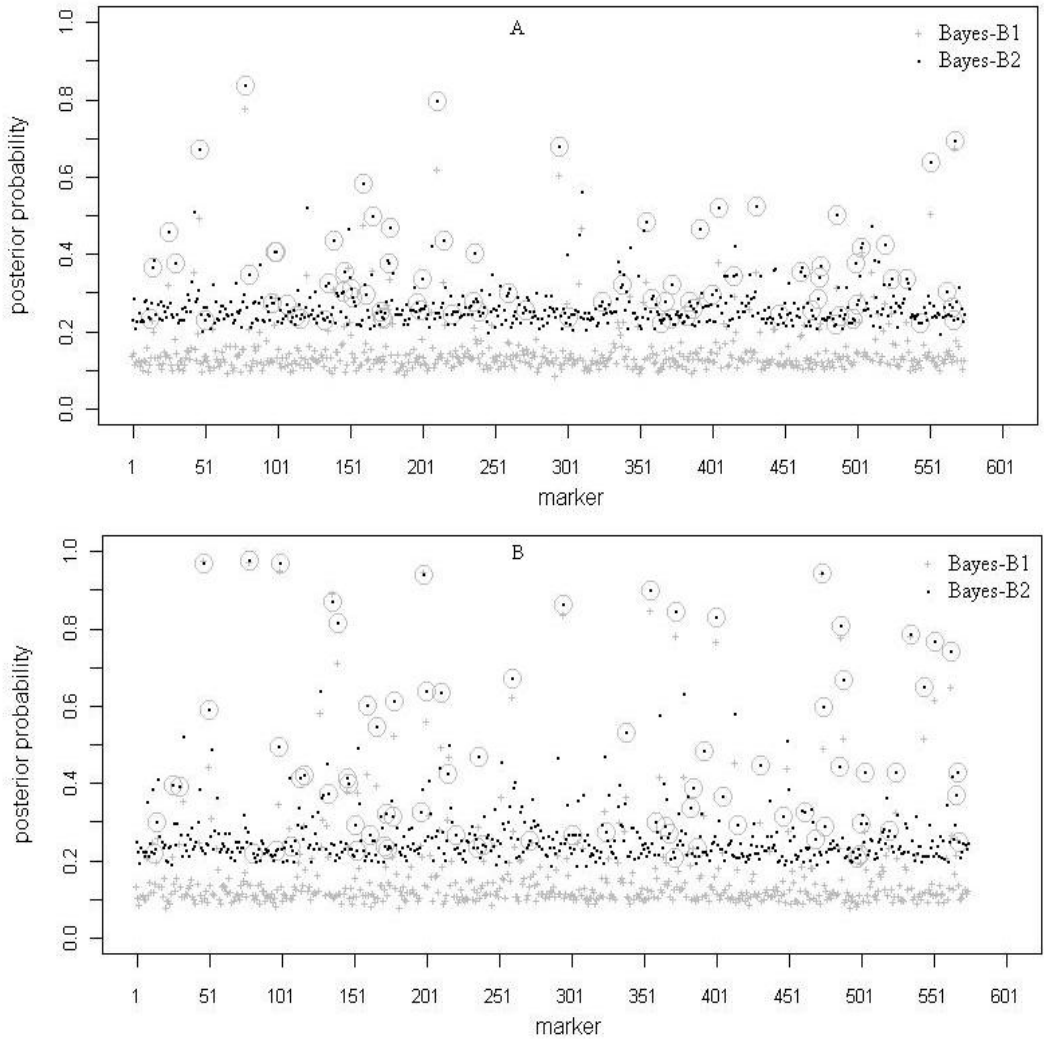
A
Bayes-B1
Bayes-B2
posterior probability
0.0
0.2
0.4
0.6
0.8
1.0
1
51
101
151
201
251
301
351
401
451
501
551
601
marker
B
Bayes-B1
Bayes-B2
posterior probability
0.0
0.2
0.4
0.6
0.8
1.0
1
51
101
151
201
251
301
351
401
451
501
551
601
marker

Figure 6

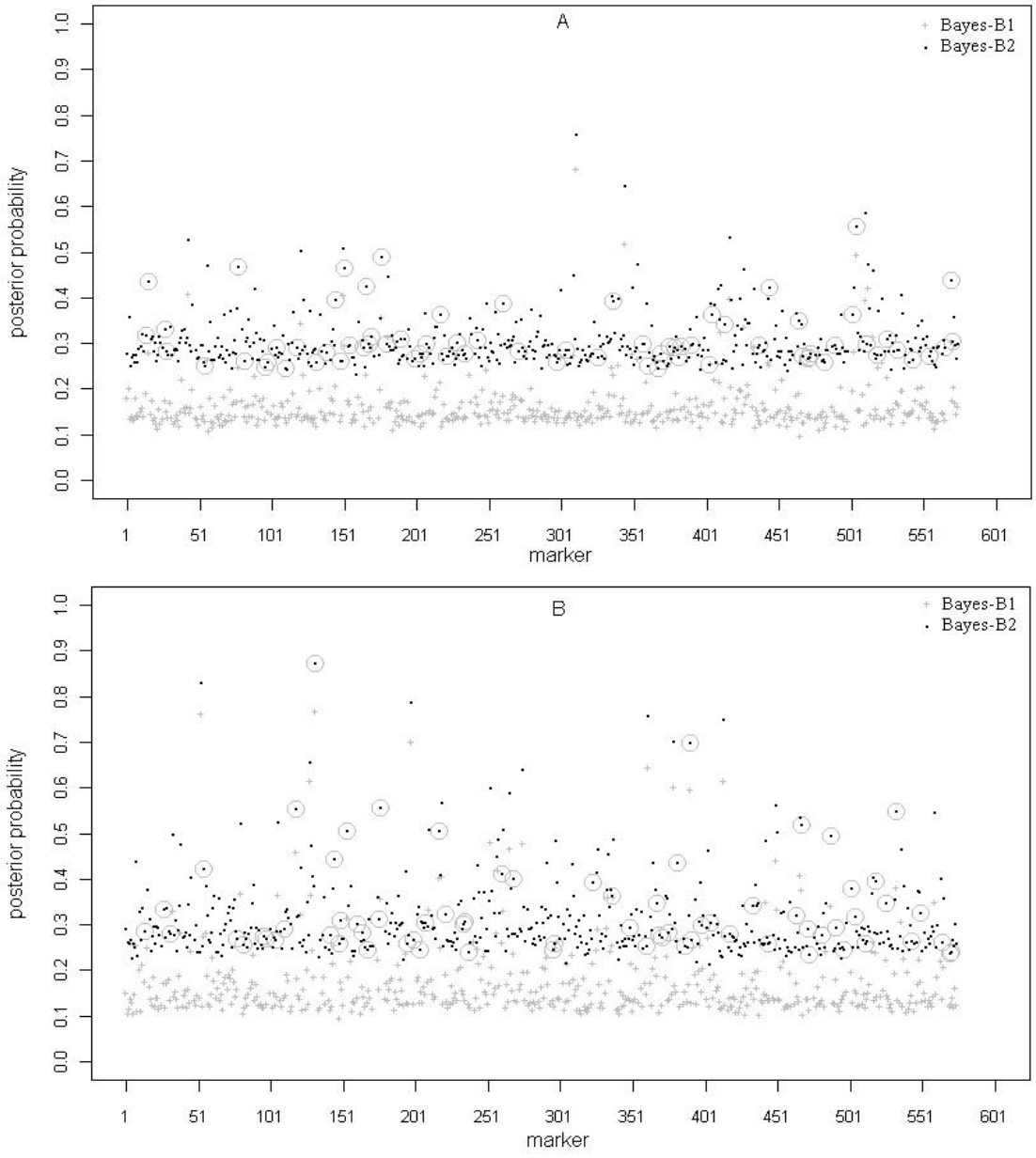
A
Bayes-B1
Bayes-B2
posterior probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1
51
101
151
201
251
301
351
401
451
501
551
601
marker
B
Bayes-B1
Bayes-B2
posterior probability
marker

Figure 7

CHAPTER V. GENERAL CONCLUSIONS

The research that is described in this thesis is primarily about integrating QTL analysis into plant breeding using Bayesian statistics. The impact of two mating designs on QTL mapping in multiple families was studied, how the cross mean and genetic variation in the cross determine the usefulness of a cross was demonstrated, and the prospects and required for genomic selection in cultivated barley were evaluated.

Chapter II carried out QTL mapping in multiple families, utilizing within-family LD and assuming markers and QTL were in linkage equilibrium (LE) at the population level. We found that the loop design was better in estimating the QTL allelic variance and position, and more powerful at a typical significance threshold than the reference design.

Chapter III investigated how the cross mean and genetic variance could be taken into account during cross selection. For inbred line development, the value of a particular cross depends not on the mean of all its progeny but on the best of its inbred progeny. We developed the theory to predict the value of the best progeny as a function of the mean of all progeny and of their standard deviation, using results from QTL analysis. The benefit of accounting for this standard deviation increased with increasing heritability and marker density used in the QTL analysis, but decreased with increasing genome size and QTL number. The benefit was also higher if only crosses among the best parents were performed, as would be typically occur in breeding programs. It was also demonstrated that the method of QTL analysis affects how well superior progeny can be predicted: including all rather than only significant markers in the calculation improved prediction, as did accounting for the uncertainty of QTL analysis. Nevertheless, we generally found little difference among crosses for the standard deviation among their progeny, such that a benefit from estimating that standard deviation occurred only in relatively few cases.

In Chapter IV, barley SNP data were used to evaluate association-based genomic selection for spring two-row barley. It was demonstrated that the current barley SNP density was high enough for genomic selection to predict progeny breeding values with reasonable accuracy. Overall, the Bayes-B approach that fitted a relatively higher proportion of markers

into regression model had more stable performance across different scenarios involving different QTL and marker numbers, size of the training dataset, and levels of linkage disequilibrium.

All genomic selection analysis methods tested in Chapter IV, except RR-BLUP, and QTL analysis methods in Chapter II and III were Bayesian methods. Therefore QTL allele effects were treated as random. The main difference among the methods was in the prior for the QTL information. In chapter II, the prior for the number of QTL was assumed to follow a Poisson distribution, the prior for the QTL variance followed a uniform distribution within a reasonable range, and the prior for the QTL position was a uniform distribution across the genome. In chapter III, a Bayesian shrinkage estimation approach was used: all markers were included in Bayesian model and the prior for the log of marker variance was uniform from negative infinity to positive infinity. In Chapter IV, Bayes-B (Meuwissen et al. 2001) was one of the three methods applied. In Bayes-B, the prior assumption is that most markers have no effect and the variance of those markers that do have effects follows a gamma distribution. This Bayes-B prior intuitively seems to approximate what we believe is true in nature. It also seemed to be the best approach, with its ability to handle high-dimension data and accommodate complex population structure. Chapter IV demonstrated Bayes-B has better performance in accounting genetic relationship than the Bayesian shrinkage analysis because the later approach tried to shrink all small effects toward zero. Therefore the accuracy would likely increase if Bayes-B approach was used in Chapter III. Chapters II and III only utilized within-family LD for QTL mapping and thus were based in linkage analysis, while Chapter IV exploited population-wide LD and thus was based in association analysis. Chapters II and IV both involved QTL mapping in multiple families. Chapter II assumed LE markers at the population level while Chapter IV used population-wide LD markers for mapping.

Yu et al (2008) demonstrated that nested association mapping (NAM) is a very powerful QTL mapping approach. The approach they used involved step-wise model selection. Chapters III and IV of this thesis showed that genomic selection works well with both simple and complicated family structure. Using a genomic selection method such as Bayes-B, which well balances its ability to capture the genetic relationship and QTL effects by LD, should

therefore be very interesting for QTL mapping in the NAM design. In order to increase the chance of identifying causal factors from thousands of variables, machine learning has also been developed to first select a subset of variables for further analysis (Long et al, 2007). Given complete sequence or very high-density markers, possibly combined with complicated epistasis and genotype by environment interaction, the number of possible explanatory variables is extremely large. It would be very difficult even with methods that are optimal for handling high dimensionality such as Bayes-B to handle this complexity. Another way to select a subset of variables could be simple t-test statistics, which are regularly used in microarray analysis to identify from tens of thousands of genes whose expression was profiled ones that have significant differential expression at a certain false discovery rate (FDR) (Nettleton 2006). A similar method could perhaps be used to eliminate useless variables in the genomic selection context. Instead of setting a low FDR, a relatively high FDR could possibly help to remove most useless but not important variables. This elimination would allow the MCMC-based random walking among the remaining variables to achieve more meaningful results and might allow us to explore high-dimension effects, such as epistasis and genotype by environment interaction.

Another alternative to handle high dimensionality might be an empirical Bayes (E-Bayes approach (Xu 2007). Xu (2007) integrated the prior knowledge for the QTL variance into the mixed model by E-Bayes approach: the maximum likelihood for the QTL variance is estimated first, and then BLUP estimates for all effects are obtained. Compared to the full Bayesian approach, the E-Bayes approach avoids possible convergence problems associated with estimating marker variance. It would be interesting to investigate the behavior of this approach in genomic selection and with other high-dimension genomic data.

With the development of statistical analyses such as those described in this dissertation and given ongoing decreases in marker costs, marker-assisted selection for complex traits in plant breeding appears to have a bright future.

References

Long N., D. Gianola, G. J. M. Rosa, et al. 2007 Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers Journal of animal breeding and genetics: 124: 377–389.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Nettleton, D. 2006 A discussion of statistical methods for design and analysis of microarray experiments for plant scientists. The Plant Cell 18: 2112-2121.

Yu J, Holland J B, M D McMullen, and E S. Buckler 2008 Genetic Design and Statistical Power of Nested Association Mapping in Maize. Genetics 178: 539-551.

Xu, S.Z., 2007 An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics 63 (2): 513-21.

APPENDIX. R CODE FOR GENOMIC SELECTION

Bayes-A (Meuwissen et al. 2001; Xu 2003; ter Braak et al. 2005)

```
## The code was partly based on the simple demonstration version of Bayes-A during The QTL Mapping, ##
MAS and Genomic Selection Workshop at Iowa State University,2007, provided by Dr. Ben Hayes.
## The code has been optimized by allowing several-chain running, burn-in,
## updating the marker effect in random order, and updating the parameters at the same time in vector
## or matrix format, etc.

## Please first go to your working directory, under working directory, create a folder named “data”
##  to store the data, a folder named “code” for the R code, and a  folder name “result” to store results.

## Define your own working directory for R here.
workingDirectory = "Z:/Shengqiang/project3/barley"
folderName=paste(workingDirectory,"/data", sep="")
setwd(folderName)  ## go into the data folder to read data.

# Data format
# Prepare your marker data as .csv files, see the example in genotype.csv file.
#     The first row is the marker loci or haplotype segments. Each other row
#     corresponds to the genotype of an individual. Since we have inbred line
#     situation, for genotype, we simply code one homozygous allele as 0, other
#     homozygous allele as 1. If you have heterozygote at one locus, you can
#     code that as 0.5 in the genotype matrix.
# Prepare your phenotype data as .csv files, see the example in phen.csv file.
# Put all your marker and phenotype data in the data folder.

# Read data
fileName=paste("genotype.csv") ## file name of Marker data
x = as.matrix(read.csv(fileName,header=T)) # Genotype
fileName=paste("phen.csv") ## file name of phenotype data
y = as.matrix(read.csv(fileName,header=T))  # phenotype
```

```
# Some useful constants, vectors, and matrices for the computation.
markerNo=ncol(x)   ## marker number or haplotype segment number.
recordNo=length(y) ## record number of phenotype
ones = array(1,c(recordNo))
ySum=t(ones)%*%y
xColSum=t(ones)%*%x
xSquareColSum=t(ones)%*%x^2
txy=t(x)%*%y
txx=t(x)%*%x
```

 # MCMC Parameters

```
iterNo=10000    # number of iterations
burnin=5000     # Generally, I think 2000 burnin is good enough.
thin=10
chainNo=2

# Storage vectors and matrices
gStore = array(0,c(1,markerNo))  # marker effect

#gVarStore = array(0,c(1,markerNo)) # marker variance
#eVarStore = array(0,c(1)) # residual variance
#muStore = array(0,c(1))   # intercept
#iterStore = array(0,c(1)) # iteration
```

 # MCMC initialization

```
g = array(0,c(markerNo,chainNo))
mu = array(mean(y),c(1,chainNo))
gVar = array(var(y)/markerNo*10,c(markerNo,chainNo))
eVar = array(var(y)/markerNo*10,c(1,chainNo))
```

Beginning of MCMC simulation.

```
for (iter in 1:iterNo) {
```

 # Step 1 Sample the g (marker effect) from a normal distribution

```
# sample marker effect in random order. This could result in better mixing between chains.
  randomOrder=sample(1:markerNo,replace=F)
  for (j in randomOrder) {
```

```
   gtemp = g
   gtemp[j, ] = 0
   mean = ( txy[j]-txx[j,]%*%gtemp-xColSum[j]*mu ) / (xSquareColSum[j]+eVar/gVar[j,])
   sd=sqrt(eVar/(xSquareColSum[j]+eVar/gVar[j,]))
   for (iChain in 1:chainNo) g[j, iChain] = rnorm(1,mean[iChain], sd[iChain])
  }
```

 # Step 2. Sample the gVar from the inverse chi square posterior

```
   # Bayes-A prior Meuwissen et al. (2001)
   # gVar =(0.002+g^2)/rchisq(markerNo*chainNo,4.012+1)

   # Xu (2003) prior
   #gVar=g^2/array(rchisq(markerNo*chainNo,1),c(markerNo,chainNo))

   # ter Braak et al. (2006) prior
   gVar=g^2/array(rchisq(markerNo*chainNo,0.998),c(markerNo,chainNo))
```

 # Step 3. Sample eVar from an inverse chi-square posterior

```
   e=matrix(rep(y,chainNo),,chainNo,byrow=F)-x%*%g-matrix(rep(mu,recordNo),recordNo,byrow=T)
   eVar = colSums(e^2)/rchisq(chainNo,recordNo-2)
```

 # Step 4. Sample the mu from a normal posterior

```
   for (iChain in 1:chainNo) mu[iChain] = rnorm(1,(ySum -
xColSum%*%g[,iChain])/recordNo,sqrt(eVar[iChain]/recordNo))

# Save results after burnin
   if (iter>burnin && (iter%%thin)==0){
     gStore =rbind(gStore,t(g))
    #gVarStore = rbind(gVarStore,t(gVar))
    #eVarStore  = rbind(eVarStore,eVar)
    #muStore = rbind(muStore,mu)
    #iterStore  = rbind(iterStore,iter)
   } # End of saving results
} # End of MCMC iteration
```

 # Results

```
gStore=gStore[-1,] # marker effects: homozygous marker effect, assuming one homozygous has no effect.
#gVarStore = gVarStore[-1,]
#eVarStore =eVarStore[-1,]
#muStore =muStore[-1,]
#iterStore  =iterStore[-1,]

# write your result to result folder
folderName=paste(workingDirectory, "/result",sep="")
setwd(folderName)

# Writing results to .csv files.
#write.csv(gStore, file="gStore.csv",row.names=F)
#write.csv(gVarStore, file="gVarStore.csv",row.names=F)
#write.csv(muStore, file="muStore.csv",row.names=F)
#write.csv(eVarStore, file="eVarStore.csv",row.names=F)

# Analysis results
gStoreMean=apply(gStore,2,mean) # Marker effect
fileName=paste("gStoreMean",".csv",sep="")
write.csv(gStoreMean,file=fileName) # Write the result.

# Return to the folder where you want to go
folderName=paste(workingDirectory, "/code",sep="")
setwd(folderName)
```

 # The end of Bayes-A

Bayes-B (Meuwissen et al. 2001)

```
## Please see previous code for Bayes-A to set up your working directory and input data files in the
## Define your own working directory for R here.
workingDirectory = "Z:/Shengqiang/project3/barley"
folderName=paste(workingDirectory,"/data", sep="")
setwd(folderName)  ## go into the data folder to read data.
```

 # Read data

```
fileName=paste("genotype.csv") ## file name of Marker data
```

```
x = as.matrix(read.csv(fileName,header=T)) # Genotype
fileName=paste("phen.csv") ## file name of phenotype data
y = as.matrix(read.csv(fileName,header=T))  # phenotype

# Define the marker or haplotype segment number that you believe would be causal alleles.
qtlNo = 80

# Scale your data for Bayes-B
sdY=sd(y)
y=sqrt(2)*y/sdY

markerNo=ncol(x)
recordNo=length(y)

# Define the function to draw the prior distribution for marker or haplotype segment variance.
bayesB_prior_dsn=function(n,pi,v,s){ # sample BayesB prior distribution in Meuwissen 2001
  # n is the sample number you want to draw from the distribution
  # pi, v,  and s are the parameters for the prior distribution. The prior distribution is that given a marker
  # or a haplotype segment, the probability that the marker has no effect is pi, and if the marker has an
  # effect, the variance of this marker will follow an inverse-chi-square (v,s) distribution.
  VgPi=rep(0,n)
  temp=(c(runif(n))>pi)
  VgPi[temp]=s/rchisq(sum(temp),v)
  return(VgPi)
}

v= 4.2339; s=0.0429; # inverse-chi-square (v,s) in Bayes-B
# Define the proportion of the markers that have no effect. For example, the following pi can be regarded
# as 80 makers have non-zero effect out of total makers.
pi= 1-qtlNo/markerNo

# MCMC Parameters
iterNo=10000    # number of iterations
burnin=5000     # Generally, I think 2000 burnin is good enough.
thin=10
chainNo=2
```

```
# Storage vectors and matricies
gStore = array(0,c(1,markerNo))
gVarStore = array(0,c(1,markerNo))
#eVarStore = array(0,c(1))
#muStore = array(0,c(1))
#iterStore = array(0,c(1))

# Initialization
g = array(0,c(markerNo,chainNo))
mu = array(mean(y),c(1,chainNo))
gVar = array(var(y)/markerNo*10,c(markerNo,chainNo))
eVar = array(var(y)/markerNo*10,c(1,chainNo))

# some useful variables in the computation.
ones = array(1,c(recordNo))
ySum=t(ones)%*%y
xColSum=t(ones)%*%x
xSquareColSum=t(ones)%*%x^2
txy=t(x)%*%y
txx=t(x)%*%x

### Beginning of MCMC simulation.
for (iter in 1:iterNo) {
  # Step 1 Sample the g (marker effect) from a normal distribution
  # sample marker effect in random order. This could result in better mixing between chains.
  # Simply put zero effect for those markers with gVar==0
 g[gVar==0]=0

 # Update the marker effect where gVar!= 0
 along=(1:length(gVar))[gVar!=0]
 gColumn=ceiling(along/nrow(g))        # column in g
 gRow=along%%nrow(g) # iChain
 gRow[gRow==0]=nrow(g)  # j-th marker
 randomOrder=sample(1:length(along),replace=F)
 for (iAlong in randomOrder){ #Update marker effects in random order.
```

```
  j=gRow[iAlong]; iChain=gColumn[iAlong]
  gtemp = g[,iChain]
  gtemp[j] = 0
  mean = ( txy[j]-txx[j,]%*%gtemp-xColSum[j]*mu[iChain] ) /
(xSquareColSum[j]+eVar[iChain]/gVar[j,iChain])
  sd = sqrt(eVar[iChain]/(xSquareColSum[j]+eVar[iChain]/gVar[j,iChain]))
  g[j, iChain] = rnorm(1, mean, sd)
 }
```

 # Step 2. Sample the gVar using Metropolis-Hasting step

```
# Sample new variance from prior.
gVarNew = array(bayesB_prior_dsn(markerNo*chainNo,pi,v,s),c(markerNo,chainNo))

# For those markers that have gVarNew==gVar, you don't need to update
# these variance because they are actually equal to 0.
# Label where gVar!= gVarNew and update gVar for those markers
along=(1:length(gVar))[gVar!=gVarNew]
uniformNumber= runif(length(along))
gColumn=ceiling(along/nrow(g))       # column in g
gRow=along%%nrow(g) # iChain
gRow[gRow==0]=nrow(g)  # j-th marker
#Preparation of some vector and matrix for updating gVar
xSquareColSumTemp = array(0,c(length(along)))
eVarTemp=array(0,c(length(along)))
yStarTemp=array(0,c(length(along)))
gVarNewTemp=array(0,c(length(along)))
gVarTemp=array(0,c(length(along)))
constantATemp=array(0,c(length(along)))
logLOld=array(0,c(length(along)))
logLNew=array(0,c(length(along)))

for (iAlong in 1:length(along)){
  j=gRow[iAlong]; iChain=gColumn[iAlong]
  gtemp = g[,iChain]
  gtemp[j] = 0
  # corrected y for other genetic factor: y-x%*%gtemp-mu[iChain]
```

```
  constantATemp[iAlong]=t(y-x%*%gtemp-mu[iChain])%*%x[,j]
  eVarTemp[iAlong]=eVar[iChain]
  gVarTemp[iAlong]=gVar[j,iChain]
  gVarNewTemp[iAlong]=gVarNew[j,iChain]
  xSquareColSumTemp[iAlong]=xSquareColSum[j]
 }
 constantATemp=constantATemp^2
 lot=(1:length(along))[gVarTemp!=0] # Locations of non-zero variance
 ## loglikelihood of the old gVar
 logLOld[lot]=-
.5*log(xSquareColSumTemp[lot]*gVarTemp[lot]/eVarTemp[lot]+1)+0.5*(constantATemp[lot]/eVarTemp[lot]
/(xSquareColSumTemp[lot]+eVarTemp[lot]/gVarTemp[lot]))
 # Locations of proposed non-zero variance
 lot=(1:length(along))[gVarNewTemp!=0]
 logLNew[lot]=-
.5*log(xSquareColSumTemp[lot]*gVarNewTemp[lot]/eVarTemp[lot]+1)+0.5*(constantATemp[lot]/eVarTemp
[lot]/(xSquareColSumTemp[lot]+eVarTemp[lot]/gVarNewTemp[lot]))
 # Metropolis-Hastings step
 ratio=pmin(exp(logLNew-logLOld),1)
 gVarTemp[uniformNumber <ratio]=gVarNewTemp[uniformNumber <ratio]
 gVar[along]=gVarTemp

 # Step 3.  Sample eVar from an inverse chi-square posterior
 e = matrix(rep(y,chainNo),,chainNo,byrow=F)- x%*%g -matrix(rep(mu,recordNo),recordNo,byrow=T)
 eVar = colSums(e^2)/rchisq(chainNo,recordNo-2)

 # Step 4.  Sample the mu from a normal posterior
 for (iChain in 1:chainNo)
 mu[iChain] = rnorm(1,(ySum -  xColSum%*%g[,iChain])/recordNo,sqrt(eVar[iChain]/recordNo))

 #  Save results after burnin
 if (iter>burnin && (iter%%thin)==0){
  gStore =rbind(gStore,t(g))
  gVarStore = rbind(gVarStore,t(gVar))
  #  eVarStore  = rbind(eVarStore,eVar)
  #  muStore = rbind(muStore,mu)
```

```
   #  iterStore  = rbind(iterStore,iter)
  } # End of saving results
}
# End of MCMC iteration

# Results
gStore=gStore[-1,] # marker effect
gVarStore = gVarStore[-1,]
#eVarStore =eVarStore[-1,]
#muStore =muStore[-1,]
#iterStore  =iterStore[-1,]
# Compute the posterior probability
p=apply(1*(gVarStore!=0),2,mean)

# Write your results to result folder
folderName=paste(workingDirectory, "/result",sep="")
setwd(folderName)

# Writing results to .csv files.
#write.csv(gStore, file="gStore.csv",row.names=F)
#write.csv(gVarStore, file="gVarStore.csv",row.names=F)
#write.csv(muStore, file="muStore.csv",row.names=F)
#write.csv(eVarStore, file="eVarStore.csv",row.names=F)

# Average the marker effects from the iterations and scale back
# corresponding to the original phenotypes scale.
gStoreMean=sdY/sqrt(2)*apply(gStore,2,mean)

#file name for the marker effect
fileName=paste("BayesB",".csv",sep="")
write.csv(gStoreMean,file=fileName)

#file name for the posterior probability
fileName=paste("BayesBprob",".csv",sep="")
write.csv(p,file=fileName)
```

```
# Back to the folder where you want
folderName=paste(upperDirectory, "barley/code",sep="")
setwd(folderName)
# The end of Bayes-B
```

References

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

ter Braak, C. J. F., M. P. Boer, and M. Bink, 2005 Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. Genetics 170:1435-1438.

Xu, S., 2003 Estimating Polygenic Effects Using Markers of the Entire Genome. Genetics 163:789-801.